Design of Armadillo Repeat Protein Scaffolds

Dissertation

zur

Erlangung der naturwissenschaftlichen Doktorwürde
(Dr. sc. nat.)

vorgelegt der

Mathematisch-naturwissenschaftlichen Fakultät

der

Universität Zürich

von

Fabio Parmeggiani

aus

Italien

Promotionskomitee

Prof. Dr. Andreas Plückthun (Leitung der Dissertation)
Prof. Dr. Markus Grütter
Prof. Dr. Donald Hilvert

Zürich, 2008

To my grandfather,
who showed me the essence of curiosity

"It is easy to add more of the same to old knowledge, but it is difficult to explain the new"

Benno Müller-Hill

Table of contents

Summary

Protein recognition is a molecular mechanism at the base of most physiological processes and used constantly in research, diagnostic and therapeutic applications. Antibodies and alternative scaffolds, developed to overcome antibody limitations, constitute nowadays binding molecules against a broad range of targets, but are all subjected to laborious selection methods for each desired target. In particular, a scaffold will be extremely valuable if it could provide specificity against peptide primary sequence and this specificity could be directly transferred to new binders, avoiding selection procedures. Therefore, protein scaffolds based on armadillo repeat proteins were designed as stable frameworks for the generation of specific peptide binders, providing a constant binding mode for peptides and proteins in extended conformation.

The generation of the framework described here was based on a consensus design strategy, involving a multiple sequence alignment of single repeat sequences and a refinement based on crystallographic data. Three types of internal repeats based on armadillo subfamilies were realized and four types of specialized capping repeats were designed to protect the hydrophobic core from solvent exposure. Designed proteins were soluble, could be produced with high yield in *Escherichia coli* and could be easily purified, but were either molten globules or present as dimers in solution. The original design was improved by applying a computational approach based on simulated annealing. This approach aimed at the stabilization of molten globule-like designed proteins by mutation of key residues in the hydrophobic core. Among the best ranking mutants, 19 proteins were analyzed and all were showing improved biophysical properties compared to the original molten globule. The computational approach allowed the conversion of a designed armadillo repeat protein with molten globule-like properties to a protein with native-like characteristics, while preserving the high expression yield and the ease of purification.

Designed Armadillo repeat proteins were used as scaffolds for the generation of libraries, randomizing the positions responsible for target interaction, and the preliminary experiments confirmed the possibility to select for specific peptide binders.

Furthermore, a second generation of armadillo repeat proteins was designed to provide an optimized geometry for the binding of the target peptides. The analysis of the repeat orientations in the natural armadillo repeat protein led to the definition of an optimal

arrangement to favor the binding of target peptides. After defining the geometrical characteristics of the designed armadillo repeat proteins, suitable sequences were generated using computational design. The designed sequences and some variants will be experimentally characterized and will lead to new improved libraries.

In the next years libraries with a variable number of repeat modules will be created to bind peptides of different size and single repeats could be selected to specifically recognize dipeptides. The long term goal is to combine pre-selected repeats to build specific binders without a need for further selection processes, improving speed and efficiency and reducing the costs of new binding molecules for research, diagnostic and therapeutic applications.

Zusammenfassung

Die spezifische Erkennung von Proteinen ist ein molekularer Mechanismus, welcher die Basis der meisten physiologischen Prozesse darstellt und daher von grosser Bedeutung für Forschung, Diagnostik und medizinische Anwendungen ist. Antikörper und alternative Proteine, die entwickelt wurden, um die Einschränkungen von Antikörpern zu überwinden, stellen spezifische Bindemoleküle gegen ein breites Spektrum von Zielstrukturen dar; jedoch ist jeweils eine arbeitsintensive Selektion für jede einzelne Struktur notwendig. Daher wäre ein generelles Bindemolekül mit Spezifität gegen eine primäre Peptidsequenz, dessen Spezifität ohne erneute Selektion direkt auf neue Bindemoleküle übertragen werden könnte, von besonderem Interesse. Da "Armadillo Repeat" Proteine sich durch eine spezifische Affinität für Peptide und Proteine in langgestreckter Konformation auszeichnen, wurden sie als Grundlage für die Entwicklung, solcher genereller Peptid Bindeproteine benutzt.

Die Entwicklung der Gerüststruktur basierte auf der Strategie des Konsensus-Designs, welche einen Vergleich verschiedener Sequenzen einzelner Repeat-Strukturen sowie eine Verfeinerung durch Berücksichtigung kristallographischer Daten beinhaltet. Letztendlich wurden drei Arten interner Repeat-Sequenzen, basierend auf verschiedenen Armadillo Unterfamilien, sowie vier Formen spezialisierter "Capping-Repeat-Strukturen" zur Abschirmung des hydrophoben Proteinkerns ausgewählt. Die entwickelten Proteine waren

löslich, in grossen Mengen in *Escherichia coli* herstellbar und einfach aufzureinigen. Allerdings lagen sie entweder als "molten globules" oder als Dimere in Lösung vor, weswegen das ursprüngliche Design durch die Anwendung eines computerunterstützten, auf Simulationen basierenden Ansatzes verbessert wurde. Dieser Ansatz zielte darauf ab, die Proteine mit "molten globule"-ähnlichen Strukturen durch Mutationen an entscheidenden Positionen innerhalb des hydrophoben Inneren des Proteins zu stabilisieren. Von denen als besonders interessant eingestuffen Mutationen wurden 19 verschiedene analysiert, die verbesserte biophysikalische Eigenschaften, verglichen mit den ursprünglichen Molekülen aufzeigten: die computerbasierte Methode ermöglichte die Umwandlung eines "Designed Amadillo Repeat Proteins" mit einer "molten globule" ähnlichen Struktur in ein Protein mit Eigenschaften, die dem des ähnlich zum urspünglichen Protein ähnelten, wobei die hohe Expression und die problemlose Aufreinigung beibehalten werden konnte.

Zusätzlich wurden die entwickelten "Armadillo Repeat" Proteine als Gerüststrukturen zur Herstellung von Proteinbibliotheken genutzt, welche randomisierte Positionen an den Interaktionsstellen zu den Zielpeptiden aufweisen. Erste Experimente bestätigten die Möglichkeit, spezifische Peptid-Binder zu selektionieren.

Des Weiteren wurde eine zweite Generation an "Armadillo Repeat Proteinen" entwickelt, um eine optimierte Geometrie für die Bindung der Zielpeptide zu erreichen. Die Analyse der Repeat-Orientierungen in einem natürlich-vorkommenden "Armadillo Repeat Protein" führte zur Identifikation einer optimalen Anordnung für die Bindung von Zielpeptiden. Nachdem die geometrischen Eigenschaften der "Armadillo Repeat Proteine" definiert worden waren, konnten passende Sequenzen mit Hilfe von computergestützten Design entwickelt werden. Diese neu-designten Proteine werden experimentell analysiert und sollten zu neuen verbesserten Bibliotheken führen.

In den nächsten Jahren werden Bibliotheken mit einer variablen Anzahl von Repeat Modulen zur Erkennung von Peptiden verschiedener Grössen entwickelt werden. Das langfristige Ziel ist die Kombination vorselektionierter Repeat Module zu spezifischen Bindeproteinen mit vorhersagbarer Spezifität, die keine weiteren Selectionprozesse unterlaufen müssen, was folglich sowohl die benötigte Zeit, als auch die Kosten der Entwickling neuer Bindeproteine für Forschung, Diagnostik und medizinischen Anwendungen reduzieren und somit die Effizienz verbessern wird.

Chapter 1
Introduction

Part I: Peptide binders

Proteins are known as one of the characteristic molecules of all living organisms. From bacteria to eukaryotes, proteins perform almost all functional and structural tasks in the cells. Maybe even more important than their function alone is the interaction with other proteins and cellular components (lipids, sugars, nucleic acids). The interaction network gives rise to metabolic pathways, supra-molecular structures, mechanism of control and replication, motility, everything that make a living cell something more complex then just the sum of its parts and able to sustain itself and propagate.

The basic principle that makes these interactions possible is the recognition of the partner molecules. As a first general approximation, the three-dimensional structure of a protein provides the means, by shape complementarity and specific interactions at the atomic level (hydrogen bonds, non-polar interactions, salt bridges), for the recognition. The specificity is often very high, allowing a distinction between similar molecules and an interaction only with a particular one.

Most proteins achieve a high selectivity for their partners by having a structure that limits the possible interactions. But in several cases a certain degree of promiscuity in interactions is preferred, like an enzyme that would be able to catalyze a reaction on several substrates. The overall structure of the protein allows the recognition of a series of partner molecules, usually quite similar; changes in the sequences around the binding site can confer new specificities to the mutants. An extreme version of this concept would lead to a protein able to recognize a wide range of completely unrelated molecules, but such a protein would lack the specificity required for its function. Nature has solved this problem in a rather elegant way, by using proteins with the same, or similar, structure but with different sequences: each protein can recognize a different target by changing only a limited number of residues. The overall structure can be considered as a scaffold in which a particular function, in this case binding to a partner, can be introduced without disrupting the preexisting organization.

As a particular function, peptide binding indicates the ability to bind, as interaction partner, a peptide, intended here to mean not only a short amino acid sequence, but also an unstructured part of a protein or a natively unfolded protein. When looking for a scaffold for peptide binding, the first step is observing what has been done in Nature, what are the most widely used scaffolds and how the peptide binding problem has been solved.

Modular interaction domains as natural peptide binders

Several scaffolds able to bind peptides were discovered over the years in prokaryotes and especially in eukaryotes, where the cellular complexity increases and a higher level of coordination is required. Indeed, several modular interaction domains, which represent the first broad group of peptide binders, are found in proteins involved in signal transduction [1; 2; 3].

The modularity derived from the ability of these domains to fold independently of the rest of the polypeptide chain, which allows them to be used as parts, or "modules", of several different proteins, performing often the same or similar function. The removal of the module usually does not affect the overall structure of the protein but abolishes the function related to that particular domain.

Modular interaction domains usually recognize exposed sites on the surface of the partner molecules with affinities in the low nanomolar to high micromolar range. Typically, such domains recognize a core sequence but flanking or noncontiguous regions can influence the strength and the specificity of the interaction. The most common modular interaction domains recognizing peptides and post-translationally modified residues are shown in Fig. 1, according to the SMART [4; 5] representation; they usually recognize, at least as core sequences, short continuous stretches of amino acids. Among them, the so called small adaptor domains, e.g. SH2, PTB, PDZ, have been extensively studied for their role in signal transduction and assembly of multi-protein complexes (Fig. 2).

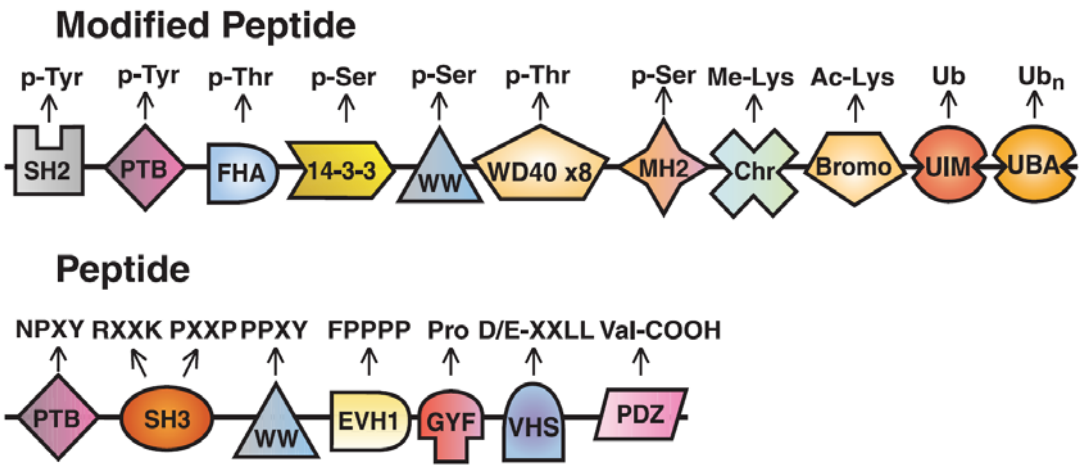


Fig. 1 The most common modular interaction domains binding peptides or modified peptides. The core sequences that are recognized are indicated. The depicted domains represent only a subset of the existing ones. The figure was adapted from Pawson and Nash [1].

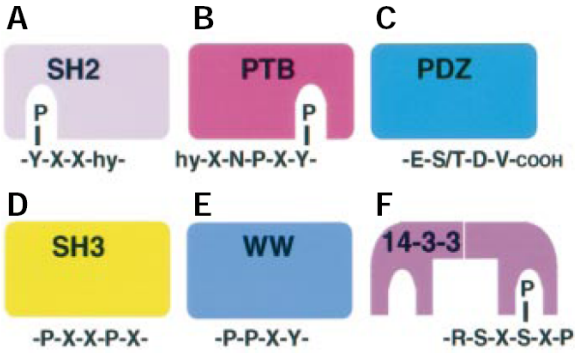


Fig. 2 Small adaptor domains used in signaling complexes. Figure adapted from Pawson and Scott [2].

SH2 domains

SH2 (src homology 2) domains [6; 7] are protein modules of approximately 100 amino acids that recognize short motifs composed of a phosphotyrosine (p-Tyr) followed C-terminally by three to five residues that determine the specificity. Proteins containing SH2 domains are involved in signaling by tyrosine kinases, both at the receptor level and during signal transduction in the cytoplasm. The SH2 domain fold is characterized by a conserved antiparallel β-sheet, containing three or four strands, surrounded by two α-helices (Fig 3a). The p-Tyr moiety is bound in a conserved positively charged pocket on the domain surface and stabilized by hydrogen bonds. The other residues of the recognition motif interact with variable SH2 residues that form the surface surrounding the p-Tyr-binding pocket.

PTB domains

The PTB (phospho-tyrosine binding) domain [6; 7] was first identified in the protein Shc and contains approximately 200 residues. PTB domains are characterized by a β-sandwich structure capped by a C-terminal helix, termed the pleckstrin homology (PH) domain "superfold" (Fig. 3b). The peptide is bound in an extended conformation between strand 5 and the helix, forming in practice an additional β-strand.

The original domain recognizes phosphopeptide motifs in which a p-Tyr is preceded by residues that form a β-turn, usually with the consensus NPXpY. Residues laying five to eight residues N-terminal of the p-Tyr also contribute to the specificity.

However, p-Tyr recognition is a characteristic of only a group of the PTB domains and the residues involved are not conserved. In contrast to SH2 domains that are devoted to p-Tyr recognition, PTB domains are, in principal, general peptide recognition modules, not relying strictly on phosphorylated peptides.

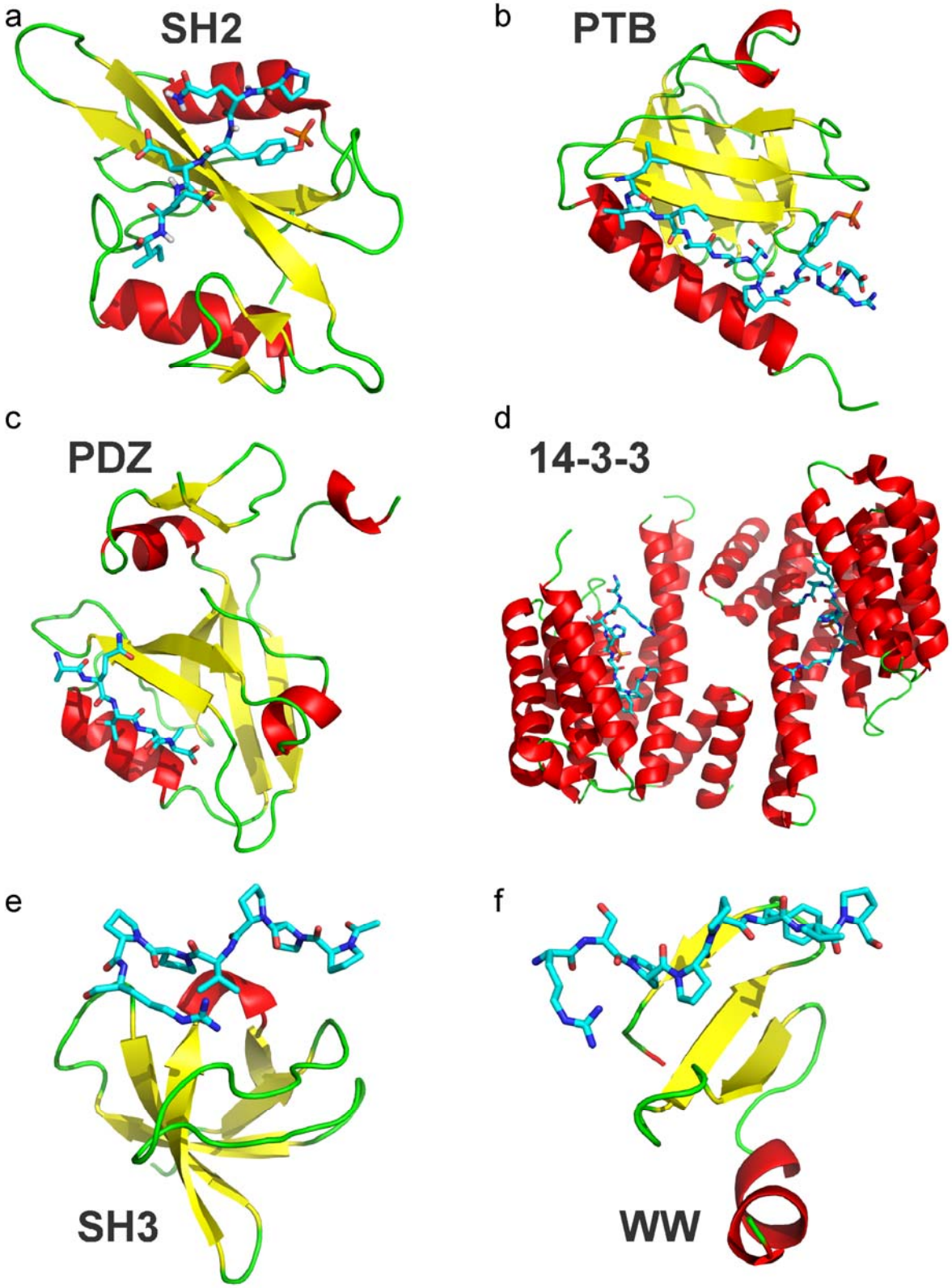


Fig. 3 Structures of complexes of small adaptor domains with their peptide target. (a) SH2 domain of src (PDB ID 1SPS) [8]. (b) PTB domain of IRS-1 (PDB ID 1IRS) [9]. (c) PDZ domain of PSD-95 (PDB ID 1BE9) [10]. (d) Structure of 14-3-3 ζ (PDB ID 1QJB) [11]. (e) SH3 domain of Sem5 (PDB ID 1SEM) [12]. (f) WW domain of dystrophin (PDB ID 1EG4) [13]. Helices are shown in red, sheets in yellow and coils in green. The target peptides are depicted in cyan. The pictures were generated with PyMol.

PDZ domains

Another well know interaction domain, which, similarly to PTB, binds the target peptide as an additional β-strand, is PDZ [14] (Fig. 3c). The name is the combination of the initials of the first discovered proteins containing the domain: postsynaptic density 95, PSD-95; discs large, Dlg; zonula occludens-1, ZO-1. The size is approximately 90 residues. PDZ domains recognize short C-terminal sequences. The last residue is hydrophobic, as exemplified by the sequence E(S/T)DV at the C-term of a group of target proteins. The specificity is conferred by residues at the –2 to –4 positions relative to the COOH-terminus of the ligand and may be regulated by phosphorylation, because the –2 residue of PDZ-binding sites is often a hydroxyamino acid. PDZ domains can be divided in several groups but together represent the third most frequent type of interaction domain, after WD40 and leucine rich repeats (LRR).

14-3-3 domains

14-3-3 domains [11; 15] derived their unusual name from the fraction number of DEAE-cellulose chromatography and its migration position in starch-gel electrophoresis used for identification of the first member in 1967. In contrast to other small adaptor domains, the members of the 14-3-3 family work as adaptor proteins in form of homo- or hetero-dimers, mainly composed of α-helices. Each 30 kDa monomer of these saddle-shaped proteins provides an extended groove for the target, often represented by consensus sequences containing phosphoserines or phosphothreonines (Fig. 3d).

Other small adaptor domains, like SH3 and WW, are devoted to the recognition of proline-rich sequences [16].

SH3 domains

SH3 (Src homology 3) domains comprise about 60 residues and typically play an assembly or regulatory role. They recognize a consensus sequence PxxP, but the specificity is provided by flanking regions. The loops between the two β sheets that form the structure contact the target in the PPII (poly-proline II) helix conformation (Fig. 3e).

WW domains

WW domains are highly compact (35 to 45 residues) binding domains, comprising an antiparallel three-stranded fold (Fig. 3f). Their binding surfaces are composed, as for SH3

domains, of a series of nearly parallel aromatic residues from which they derived the name. The binding site is, however, smaller and recognize only a core xP sequence, relying heavily on target flanking sequences to increase affinity and specificity.

All these small adaptor domains, and many others found less frequently, are generally characterized by a low affinity to the core sequence they recognize and additional interactions need to be provided. Target flanking residues can be recognized by another binding surface of the domain or by other domains, of same or different type, on the same polypeptide, creating an avidity effect. However, the recognition of short sequences almost precludes the achievement of high affinities and even the combination of several modules does not grant the binding of a continuous stretch of amino acids. Small adaptor domain binding depends, often completely, on a particular sequence feature they recognize, usually a phosphorylated amino acid, a stretch of proline residues or a free C-terminus. These domains represent the solutions to specific recognition problems, but seem to lack the possibility to be general peptide binding scaffolds, which can be found in other types of peptide binding proteins.

Peptide recognition by major histocompatibility complexes

A second class of peptide binding proteins is constituted by the major histocompatibility complexes (MHC I and MHC II) [17; 18]. These membrane-associated proteins of approximately 40 kDa are key players in the mammalian immune response and in the establishment of self tolerance, preventing the immune system to react against polypeptides normally present in the body.

Both MHC class I and class II molecules recognize peptides but differ in the kinds of peptides bound and the intracellular processing required to associate the peptide with the protein. MHC I proteins bind to short extended peptides (7-10 residues) produced intracellularly, such as viral peptides, and interact with $CD8^+$ T cells via the T-cell receptor (TCR), inducing a cytotoxic response. Class II proteins can bind, in a polyproline type II conformation, longer extracellularly derived peptides (up to ~ 20 residues) that have undergone intracellular processing. These protein-peptide complexes activate, by binding to TCR, $CD4^+$ T helper cells, which release the cytokines that play a crucial role in antibody production, cell mediated response, and other immune responses.

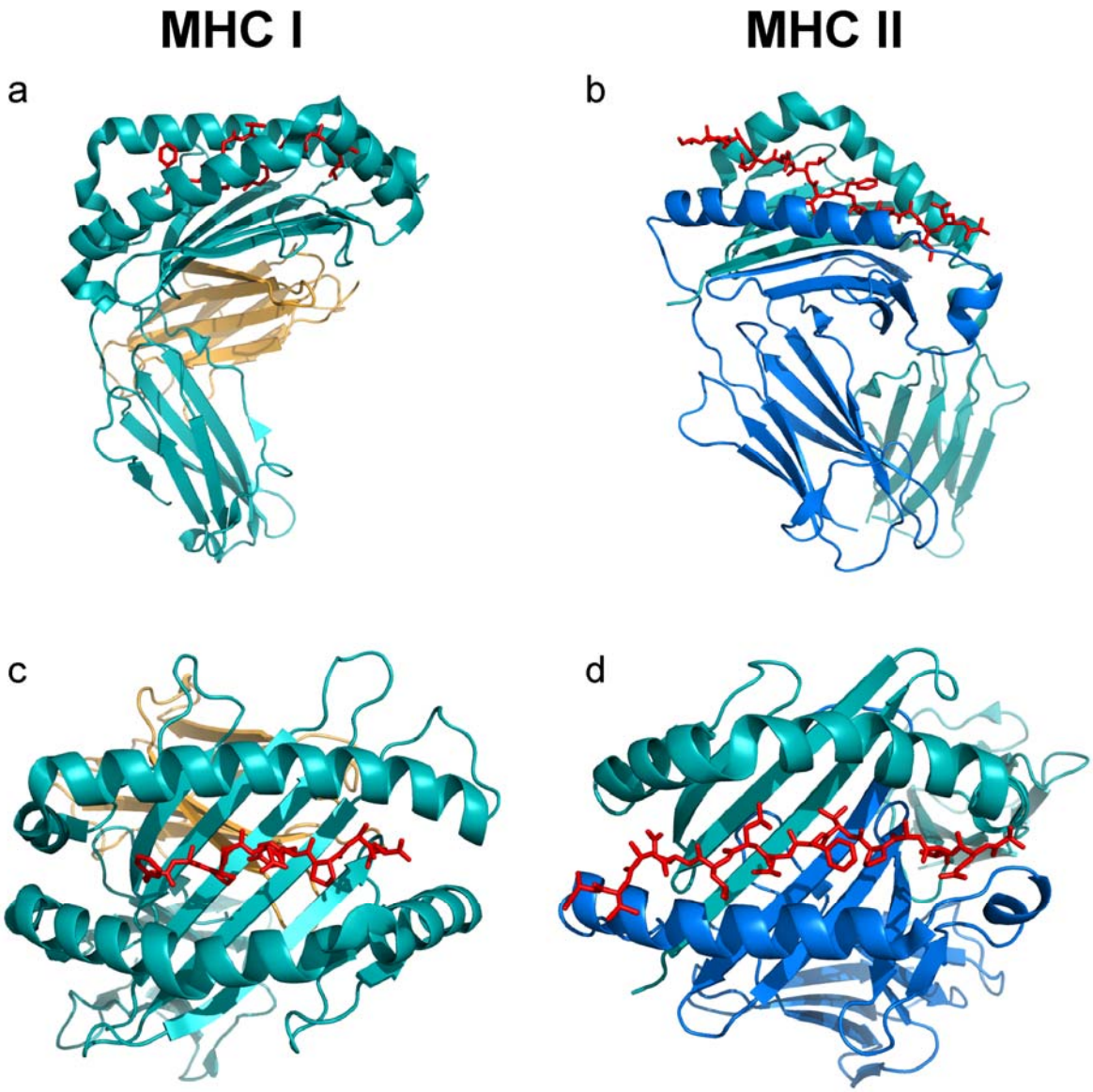


Fig. 4 Structures of MHC complexes. Side (a) and top view (c) of MHC I (PDB ID 1FZK) [19]: the α chain is depicted in cyan, β2 microglobulin in orange and the target peptide in red. On the right side, side (b) and top view (d) of MHC II (PDB ID 1ES0) [20]: α and β chains are shown in cyan and blue, the bound peptide in red. The pictures were generated with PyMol.

Structurally, both classes are characterized by a peptide binding cleft formed by an antiparallel β-sheet and two α-helices (Fig. 4). The peptide is positioned between the two helices and contacts the β-sheet at the bottom of the cleft. In MHCI the helices are almost in contact at the extremities, narrowing the binding site and limiting the access to shorter peptides. In MHCII the cleft is open and longer peptide can be accommodated, with N- and C-terminal parts outside of the cleft.

MHC I and II molecules are polymorphic and exhibit allele-specific preferences in peptide binding: residues at specific position bind well defined pockets in different alleles and the free target side chains are available for interaction with TCRs [17; 18].

Repeat proteins targeting peptides

Repeat proteins are a broad group of molecules containing at least one domain described, at the structural level, as a combination of simple modules or repeats formed generally by few secondary structure elements. Each domain is formed by one type of repeats, arranged consecutively in the sequence, which determine the overall topology and structure [21; 22]. Repeat proteins are characterized by, and limited to, short range interaction between adjacent repeats. Solenoid structures are the most common [23; 24], but a few repeat proteins possess a circular, or "closed", structure where the first and last repeat interact with each other, similarly to any other pair of adjacent repeats. In contrast to modular interaction domains, the single repeats are not able to fold independently and they assume the correct conformation only in the context of the full length domain.

At the sequence level, repeat proteins are usually characterized by repeat consensus sequences and profiles that can also be used for repeat search and classification. However, the level of similarity between repeats of the same type can be as low as approximately 10% for HEAT repeat proteins [25; 26] or almost non-detectable. The limit case is the beta propeller structure, obtained with two completely unrelated types of repeats, WD40 and kelch, probably as result of convergent evolution [27]. It is important to notice, especially for engineering repeat proteins, that a higher similarity between repeats in the same protein gives rise to a more regular structure.

Repeat proteins include leucine rich repeats (LRR), ankyrin, HEAT, hexapeptide repeat proteins and many more, and they are generally involved in protein-protein recognition. Among them Armadillo, TPR and beta propeller proteins are known as peptide binding molecules, even though not all members of these families bind peptides.

Armadillo repeat proteins

Armadillo repeat proteins [25; 28] are characterized by repeats of approximately 42 residues arranged in three α helices. The overall structure is a right-handed superhelix formed by 4 to 12 repeats. The target peptide binds in a groove on the surface in extended conformation (Fig. 5a). The name of the family corresponds to the first identified gene containing these sequences, the *Drosophila melanogaster* homolog of β-catenin, armadillo.

Tetratricopeptide repeat proteins (TPR)

TPR proteins [29; 30] form a similar right-handed super helix (topologically identical to the structure of 14-3-3 proteins), based on repeats of 34 residues on average, as indicated by the name. The repeats fold in a pair of α-helices called A and B. The inner concave surface of the super helix, formed by A helices, is used as peptide binding site, but also other surfaces are involved in protein-protein interaction (Fig. 5b).

Beta propellers

Beta propellers are the most common "closed structures" observed among repeat proteins. They are formed by a variable number of four stranded β-sheets (the so called "blades") corresponding to the repeat units. Target peptides, often carrying post translational modifications, are usually positioned on the central axis of the molecules and recognized by the loops between the β strands (Fig. 5c). Beta propellers are also involved in protein-protein interactions, but binding usually takes place at the equatorial periphery of the proteins. Beta propellers can contain WD40 or kelch repeats, but even other sequences appear to be able to assume the same fold. A WD40 repeat comprises a 44–60-residue sequence that typically contains the GH dipeptide 11–24 residues from its N-terminus and the WD dipeptide at the C-terminus [31]. A kelch repeat is 44 to 56 amino acids long and contains a GG motif, in addition to four hydrophobic residues immediately preceding GG, and conserved Y and W residues located C-terminally of the GG dipeptide. The spacing between Y and W is also highly conserved. The name "kelch" derives from the characteristic phenotype observed in the Drosophila ORF1 mutant [32].

Repeat proteins fulfill a role different from small adaptor domain. The larger size and binding interface provide higher affinity but also the possibility of binding more than one target, at the same time or in an exclusive manner, making them key players in protein

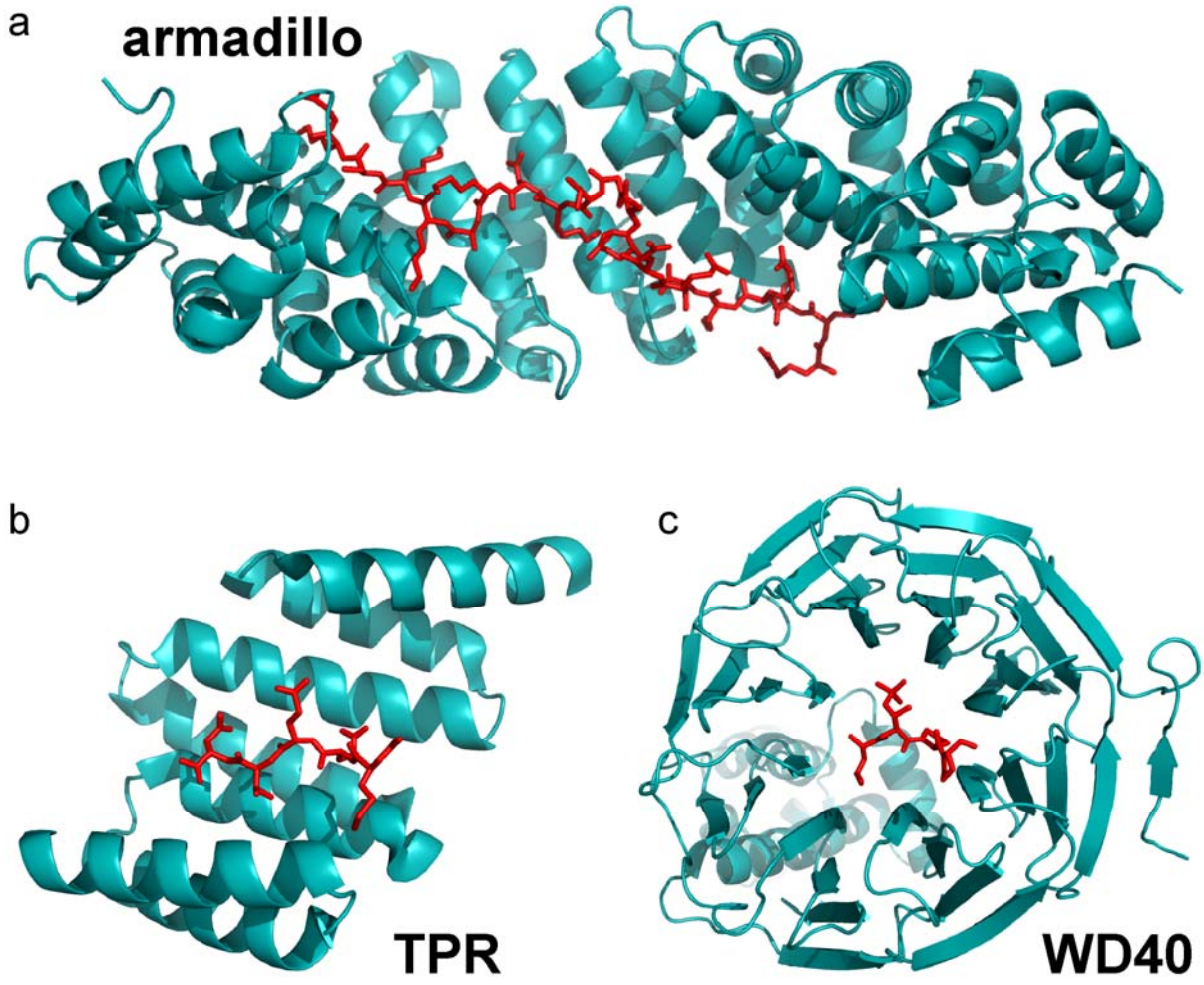


Fig. 5 Peptide binding repeat protein families. (a) Armadillo repeat protein importin-α (PDB ID 1EJY) [33], (b) TPR protein Hop (PDB ID 1ELR) [34], (c) WD40 ubiquitin ligase Cdc4 (PDB ID 1NEX) [35]. The bound peptide is shown in red.

networks. The ability of members of the same family to bind completely unrelated targets confirms that the fold has been employed several times during evolution for different functions and that its overall properties have been probably only marginally affected by the modification occurred to provide the new functionalities.

In terms of peptide binding repeat proteins are more versatile than small adaptor domains and do not have strict sequence requirements. Over the last years it became clear that this versatility can represent an advantage when it is necessary to provide different specificity by keeping the same overall structure. The E3 ubiquitin ligase complexes, where the target binding domain is often a beta propeller or a LRR [36], and the immune system of sea lamprey (*Petromyzon marinus*) based on LRR [37] are only the most notable examples, and more is

probably yet to come, by looking at the discovery of new functions for repeat proteins in plants [38; 39; 40].

Peptide binding by antibodies

Antibodies represent the prototypical binding proteins, for historical reasons, for the role in the immune system and for their ability to be selected against specific targets. The combination of two heavy and two light chains, composed of constant and variable immunoglobulin domains, produce a Y shaped quaternary structure that provides two symmetrical target binding sites (Fig. 6a).

The immunoglobulin fold [41] (called also "β-sandwich") is characterized by two antiparallel β-sheets facing each other and usually stabilized by disulfide bridges (Fig. 6b and 6c). The loops connecting the β strands are responsible for binding. However, only the N-terminal (variable) domains of the antibody chains are involved in target recognition and the corresponding loops are known as complementary determining regions (CDR). The immunoglobulin fold is also present in several other proteins involved in molecular recognition (T-cell receptors and the more distantly related integrins and fibronectin) but their main characteristics correspond to what has been observed in antibodies.

Antibodies deserve a particular analysis among peptide binders and binders in general. The CDRs are quite variable in length and sequence and determine the size and shape of the recognition surfaces. Despite the variability, the type of target correlates with the shape of the binding surface. Flat surfaces are involved in protein-protein recognition and deep pockets are the hallmark of peptide and especially hapten binding antibodies [42; 43; 44; 45] (Fig. 6d).

Immunization and selection techniques have been extensively and successfully used for development of specific antibodies [46] that represent nowadays probably the most widely studied and exploited proteins, both in basic and applied research. However, the variability that allows the selection of different specific antibody drastically reduces the possibilities of prediction or design of a binding interface. Even when observing complexes with simple targets like peptides, the target backbone assumes different position in all the complexes, the positions of the CDRs involved in binding vary and no common binding mode is detected (Fig. 7a).

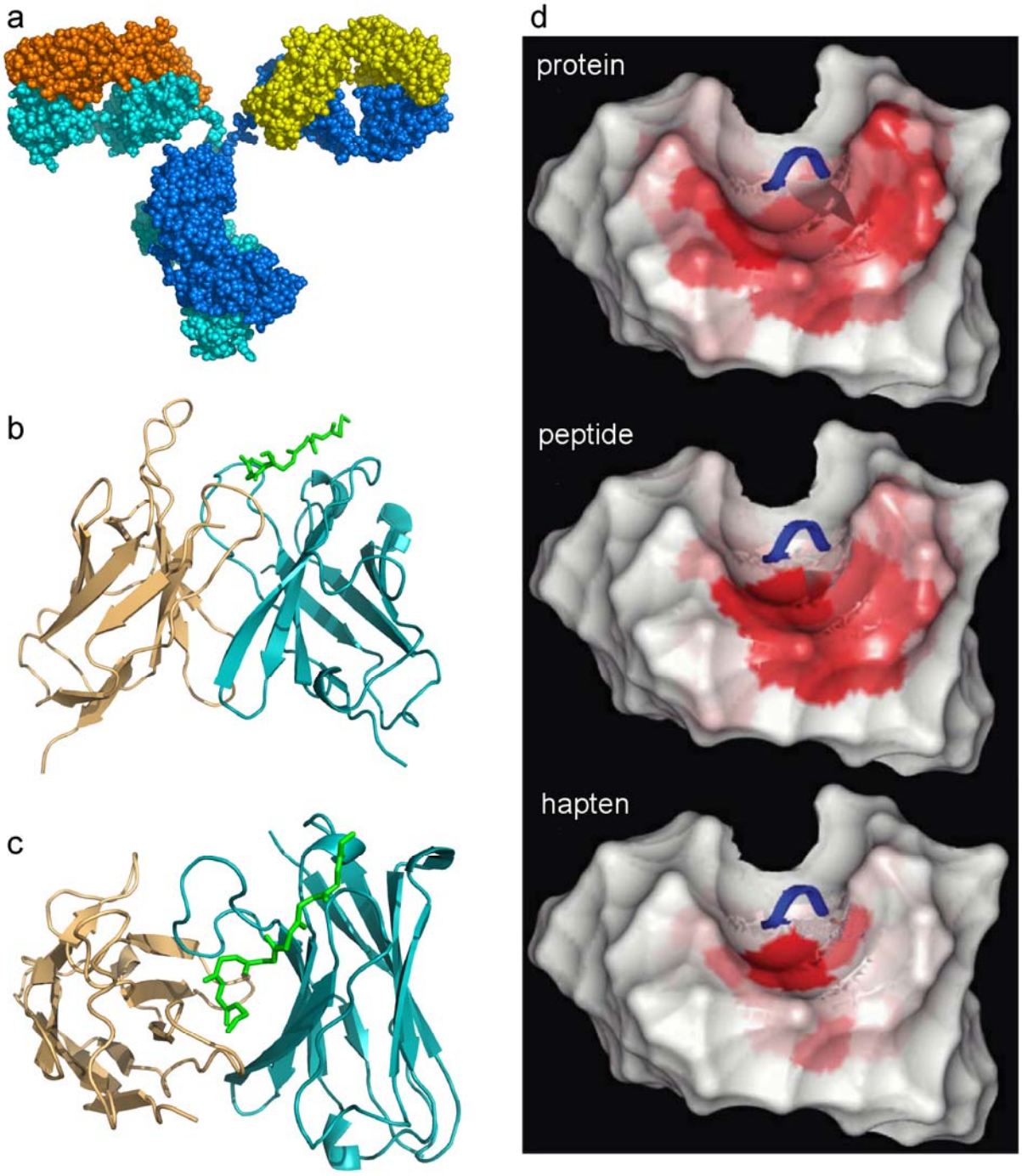


Fig. 6 Peptide binding antibodies. (a) Structure of an antibody (PDB ID 1IGT) [47]. Heavy chains are depicted in cyan and blue, light chains in yellow and orange. Side (b) and top view (c) of variable domains of an antibody (PDB ID 1HIM) [48] bound to hemagglutinin peptide (only backbone, in green). The light chain is shown in dark yellow and the heavy chain in cyan. (d) Representation of the surfaces (in red) involved in binding of proteins, peptides and haptens. The blue ribbon indicates the approximated position of the complementarity determining region (CDR) H3 which was excluded from the calculation of the binding surface. The size of the binding site and the number of residues involved are reduced according to the target size. Adapted from Almagro [42].

Alternative scaffolds: a new approach to peptide binding

The research on antibodies and the development of new related technologies brought the antibodies to the market, not only for research but also for diagnostic and therapeutics. However, antibodies proved to be not ideal for certain application and large scale production, due to their size, quaternary structure, presence of cysteines. New molecules have been generated to overcome these limitations, based on engineered antibody fragments [46] or other natural folds used for binding of protein or other molecules. The biophysical properties and the ease of production of such alternative scaffolds represent usually a significant improvement compared to the original proteins [49].

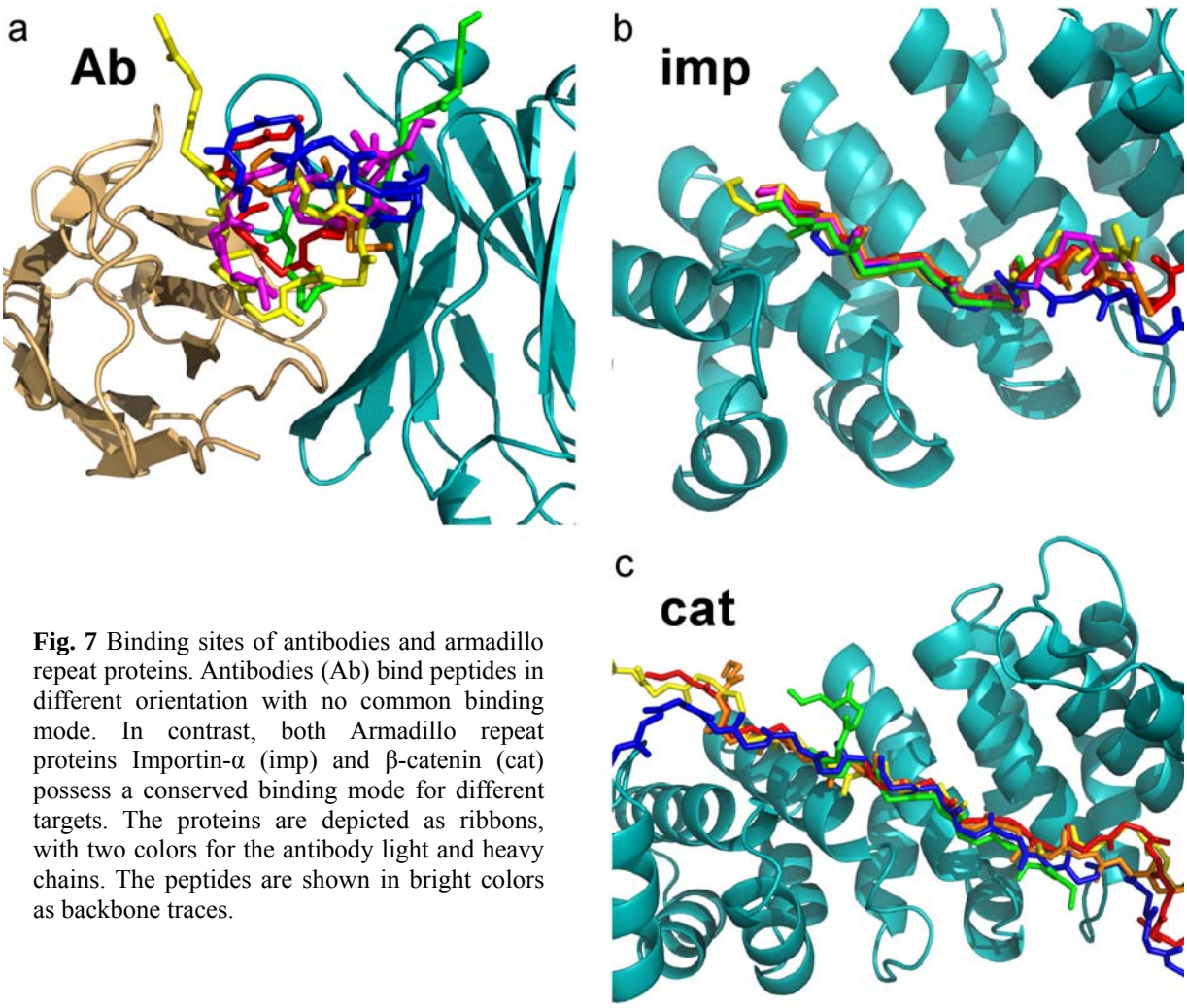


Fig. 7 Binding sites of antibodies and armadillo repeat proteins. Antibodies (Ab) bind peptides in different orientation with no common binding mode. In contrast, both Armadillo repeat proteins Importin-α (imp) and β-catenin (cat) possess a conserved binding mode for different targets. The proteins are depicted as ribbons, with two colors for the antibody light and heavy chains. The peptides are shown in bright colors as backbone traces.

Depending on the scaffold, target recognition rely on loop regions, as for antibodies, or on secondary structure elements [50]. In both cases the interaction with the target can vary quite substantially and is difficult to predict, despite the advancement in surface design. However, a peptide has a lower complexity and its binding can be theoretically more easily predicted. A constant binding mode that positions the target backbone in a fixed conformation, independently of the sequence, is the main requirement for this prediction (Fig. 7b and 7c).

Among the natural scaffold armadillo repeat proteins possess this feature, since they bind target peptides in an extended conformation and a conserved orientation. The repeat structure allows at the same time the engineering of single repeats and the combination of them. And even more important, it is possible to establish a connection between a repeat and the portion of the target it bounds. Their structural organization can give the possibility to design specific repeats able to bind short peptide fragments (2 to 3 amino acids) and combine them to generate new binders without performing additional selections.

Approaches for the engineering and stabilization of proteins and especially repeat proteins have been developed and applied to armadillo sequences to generate a new type scaffold and the first one specifically designed for peptide binding.

Part II: Principles and strategies of protein design

It is commonly accepted that that the structure and the function of a protein are the result of an evolutionary process, similar to what happens at the level of organisms. Evolution is known to work under conditions of selective pressure, in which only the fittests survive. However, the characteristics of the individuals who survive are not necessarily the best that can be achieved but only as good as to overcome the selective pressure.

The same concept applies to proteins: most of them are, for example, only marginally stable from the thermodynamic point of view, because a high stability is not required for their function. In contrast, when proteins are produced for research purposes or commercial applications, cost, ease of production, yield, stability, protease resistance, etc., are highly regarded as valuable properties which proteins, usually, have not been selected for. In terms of library generation, a scaffold with high stability allows to make mutations that create novel properties, which would be too harmful for the protein without such increased stability [51].

When it is not possible to manipulate the environmental conditions to improve protein

performance, it is, however, conceivable to modify the proteins to achieve the desired properties. The field of protein engineering is based on this assumption and focus on the modification, in general following structural hints, of known proteins. Sometimes, the protein with the desired properties is not available, simply does not exist (or it has not been discovered yet) or the modification of a known molecule to fulfill our requirements is so extensive or so complex that the final results are unpredictable. In these cases a protein cannot be simply modified but has to be designed and generated, usually using existing proteins as starting point.

In the contest of this work, the term "protein design" indicates a broad group of approaches used to design new proteins, ranging from methods based on information from multiple sequence alignments (consensus design) to algorithms for the repacking of the side chains (computational approaches).

In contrast, the traditional structure-based approaches cover usually only local changes, while protein design methods tend to provide a global modification. The introduction of a few mutations based on crystallographic data has been used, in particular, for increasing thermodynamic stability [52]. As drawback, structure-based modifications can introduce more often local negative effects, but the protein design methods can more easily compromise the overall structure of the final protein.

As an alternative to protein design, directed evolution methods, based on predefined or randomly generated libraries, have been successfully used to develop proteins with new functions or improved properties, relying on high selective pressure and the survival of the fittest molecules.

The use of directed evolution is, however, not only alternative but complementary to protein design: a lead or scaffold molecule can be created and then subjected to rounds of selections. In fact, despite all the effort in the prediction and generation of new protein sequences, these selection methods can almost always provide a good, if not better, unpredicted alternative solution.

Consensus design

Consensus design can be defined as a sequence-based and evolution-related method. Homologues proteins are seen as the result of appearance of mutations, considered as

randomly occurring events, and natural selection. Most of the mutations are neutral, while several can provide an advantage, or more often a disadvantage. They are then propagated in a population as long as their combined effects do not negatively affect the individuals. When the sequences of homologous proteins are compared, is it possible to roughly identify a set of conserved residues across the whole family and a group with high level of variability. If the mutations are considered as random events, the more frequent is a residue, the higher the probability of a positive effect: once the mutation arises, it is kept in the family, increasing the frequency. On the other hand, residues with negative effects would in fact not appear in several sequences, especially when comparing distantly related homologues. They would be more often the result of independent mutations and not conserved or at least not subjected to a strong selective pressure. Based on these assumptions, the consensus sequence obtained will contain the most frequent residues [53] . An example of consensus sequence of armadillo repeats is given in Fig. 8.

Consensus design is a relatively simple method and it does not discriminate the residues based on their roles (e.g. stability, catalytic activity, etc.). It does not require any knowledge concerning function or structure, even though this additional information has often been proven to be useful. A limiting case for a consensus design approach is the presence of alternative combinations or residues which are mutually exclusive. Some of the most frequent residues at different positions can be incompatible and the consensus sequence may present several problems, from misfolding to impaired catalytic activity. Even though a covariance analysis can provide insights into correlated residues [54], a final indication of the solution is often provided only by high resolution structural data.

A way to strengthen a consensus designed sequence is then to combine it with information from structure-based approaches.

Structure-based approaches

The term structure-based approach indicates here a group of methods used traditionally to identify advantageous point mutations by detailed analysis of high resolution structures. Most of the underlying hypotheses and solutions have been developed in the attempt to increase protein stability [2; 5] but the concepts can be extended to any other case.

Entropic stabilization refers to the approaches to decrease the free energy of the unfolded

Fig. 8 Example of a consensus sequence. In the upper section, part of the sequence alignment of armadillo repeats is shown. On the left is indicated the UniProt entry of the protein to which the repeat belongs. Amino acids are represented in a single letter code and with the color corresponding to their properties (blue, basic; red, acidic; dark green, polar; yellow, hydrophobic; orange, aromatic; cyan, histidine; light green, glycine); black positions are deletions. The lower part shows the three most frequent residues at each position obtained in the complete alignment, colored according to their relative frequency.

state and hence decreases the energy difference between folded and unfolded state. The entropy contribution in the unfolded state is mitigated by a reduction of the flexibility of the polypeptide chain: introduction of disulfide bridges, shortening of solvent exposed loops and mutated residues with reduced side chain entropy (e.g. proline) are all possible strategies to achieve it. The structure of the target protein is usually the discriminating factor to determine which one, if any at all, is suitable.

Enthalpic stabilization is based on formation of new interactions (hydrophobic, hydrogen bonds or charge effects) that decrease the enthalpy and hence the free energy of the native state, increasing the energy difference from the unfolded state. Hydrogen bonds and charge interactions are generally local changes, even though they have an overall effect on the protein, and can be relatively easily suggested by analysis of the structure. When the modifications affect the packing of the core, the effect of mutations can be predicted easily only if the region is relatively underpacked, like cavity that can be filled by longer side chains. In a well packed core, however, mutations can disrupt the packing in unpredicted ways, causing the rearrangements of the neighboring side chains and propagating the effect through the whole protein. In similar cases, a computational approach able to take into account alternative cores is probably best suited.

Secondary structure propensity is often taken into account when introducing stabilizing mutations. It is based on the assumption that residues have different backbone-dependent constraints as a function of the main chain secondary structure. However, despite the studies of propensities in peptides and proteins, the real influence is difficult to predict because other effects (e.g. shape complementarity, van der Wals interactions, hydrogen bonds) can overcome the propensity effect.

A last common approach is the replacement of exposed hydrophobic residues on the protein surface with polar amino acids. These mutations do not usually affect the thermodynamic stability or the solubility, but can reduce the aggregation rate, especially during the folding process.

Structure-based approaches are generally suitable when only few mutations are introduced. For complex tasks, like core repacking or resurfacing of a molecule, more powerful computational approaches able to consider global effects are required.

Computational approaches

The term computational approach used here includes methods of protein design based on algorithms for the identification of sequences able to satisfy the structural requirement of the desired type of fold. These methods require a backbone trace as starting point that could come from a real structure or being just a model or an ensemble of structures. The side chains are then built to provide a correct packing of the hydrophobic core and a polar surface providing solubility or even the ability to recognize a partner.

Every design program is composed of two parts: an energy function that considers the effect of van der Waals interaction, hydrogen bonds, solvation and charge interactions, and a search algorithm that explores the sequence and rotamer space to find solutions that minimize the value of the energy function. Search algorithms can be divided in stochastic and deterministic, compared by Voigt et al. [55]. Stochastic methods like Monte Carlo and genetic algorithms perform a random search in the sequence space. In principle, every change is accepted if it reduces the value of the energy function, until a minimum is reached. Genetic algorithms are based on the combination of favorable mutations or stretch of sequences. They allow the protein sequences to escape more easily from local energy minima than in the Monte Carlo methods by mixing sequence fragments, but can miss favorable sequences if residues at several positions are tightly coupled. Deterministic methods account for a systematic sequence space search. Self consistent mean field searches the sequence space using a statistical description (mean field) of the rotamers and their interactions and returns the relative preference of each amino acid at each sequence position. The complexity of the search problem is reduced and the algorithm performs well also with large proteins, but the result of the convergence (self consistency) does not necessarily correspond to the global energy minimum. In the case of Dead End Elimination the unfavorable combination of rotamers and sequences are progressively eliminated, reducing the sequence space until a single solution is found. If the algorithm converges, the result is the global energy minimum. However, for large systems convergence is not achievable in a reasonable time and the other methods are more efficient.

Given the complexity of interactions involved, computational tools have been developed allowing not only repacking, but also the *de novo* design of hydrophobic cores and entire sequences[56]. The available algorithms are able to predict sequences leading to folded proteins starting from a given backbone [57; 58; 59; 60; 61; 62]. However, even with the use of restricted

rotamer libraries [63], the application of these algorithms is still limited by the computational load, which is related to the number of variable positions and side chain rotamers considered for each residue. For these reasons, the sequences designed so far belong mainly to small proteins (up to approximately 100 residues), with the exception of an α/β barrel of more than 200 residues, designed by taking advantage of the high level of internal symmetry [64].

Design of repeat proteins

Repeat proteins, as mentioned in chapter 1, are among the most common families involved in protein-protein recognition. The interest in obtaining new stable binding molecules promoted the research in the direction of designed scaffolds derived from repeat proteins.

As postulated by Forrer et al. [65] (Fig. 9), consensus design approaches can be extremely efficient with repeat proteins, taking advantage of the large data set of repeats. The single repeats, and not the whole proteins, are, in fact, the sequences used for the definition of a consensus. The creation of building blocks allows the construction of proteins of different length, modulating the size of the binding site by choosing the number of repeats. The approach led to the creation of designed leucine rich repeat (LRR) proteins [66], tetratrico peptide repeat (TPR) proteins [67], ankyrin repeat proteins [18; 19] and beta propellers [68].

In combination with consensus design, a structure-based refinement was necessary to ensure that the repeats possess compatible interfaces allowing the folding into the desired structures. The proteins forming a solenoid structure required the addition of capping repeats at the N- and C-termini to seal the hydrophobic core, prevent aggregation and increase the solubility [21].

The designed proteins were expressed in higher yields and were thermodynamically more stable than the natural counterparts, allowing, in the case of designed ankyrin repeats, the generation of a library used for selection of specific binders [69].

A similar approach was used for designed armadillo repeat proteins and it is described in detail in Chapter 2. An alternative computational design, generated using the ROSETTA software [70], is presented in Chapter 5.

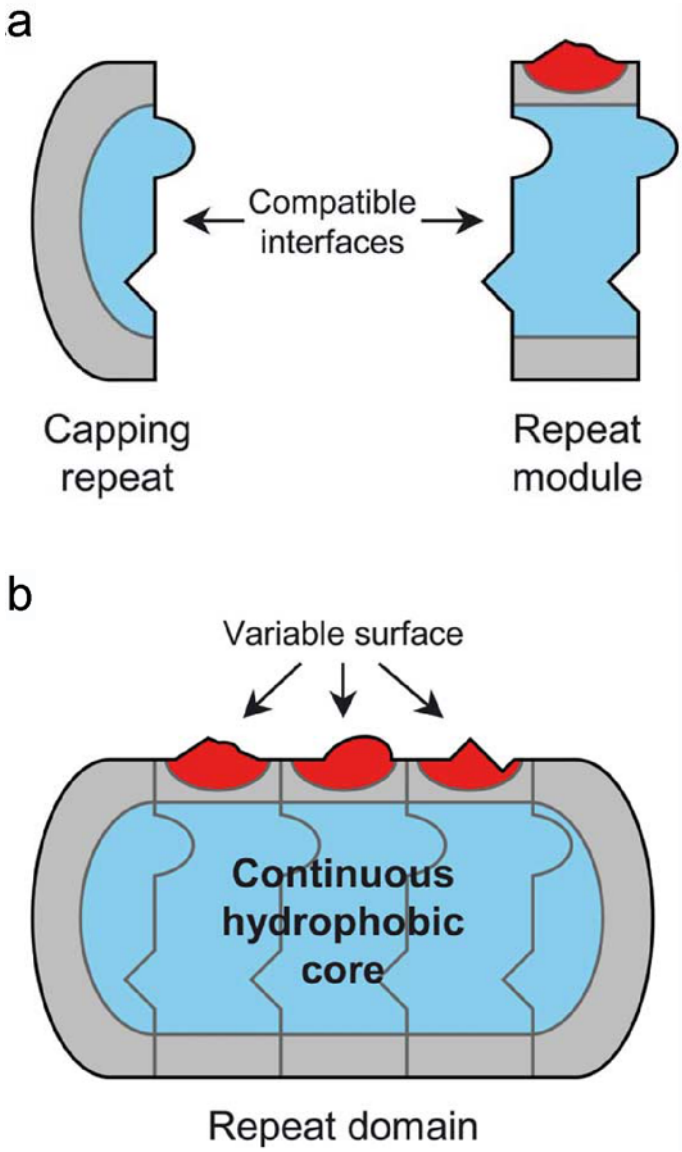


Fig. 9 Schematic representation of designed repeat proteins. (a) Compatible interfaces allow the stacking of repeat modules. (b) The repeat modules form acontinuous hydrophobic core, which is protected on both sides by capping repeats. In red is represented the variable surface responsible for binding and that is eventually randomized in a library. The hydrophobic core is shown in blue and the polar surface of the protein in gray. Figure adapted from Forrer et al. [65].

Part III: Protein libraries

A library, in general terms, can be considered as a collection of analogous objects. In biology, libraries of natural or synthetic chemical compounds are often used for drug discovery, collections of plasmids or bacterial strains are maintained, genome fragments and single genes are stored in appropriate vectors.

The main purpose of a library is to provide a collection of molecules among which at least one, more often several, possess some desired property, in terms of sequence, biophysical characteristics, affinity, etc. Therefore, a method for analyzing libraries is required, usually in

form of selection, in which the different members compete, or screening, in which all the molecules are tested separately and the results are compared. In practice, even though a library can comprise only few members, almost exclusively large library are used because of the higher chance to find a molecule with the desired properties. The ability to find it depends both on the quality of the library and on the efficiency of the selection or screening method.

Protein libraries used for protein engineering and direct evolution often comprise several variants originated from a parental protein, either natural or designed. The variants differ in sequence from the original protein but their three-dimensional structure is supposed to be conserved, at least for most of the members. They are then tested for the desired properties with appropriate *in vitro* or *in vivo* screenings or selections.

The most convenient way to store protein libraries is, however, in form of genetic material (DNA usually because of its stability) coding for the proteins. Libraries in DNA form can be easily amplified, modified and handled; the proteins are, then, produced during the selection or screening procedure.

The handling of protein libraries as DNA molecules allows the use of a series of molecular biology techniques to modify the sequence and to generate the library in the first place. Generation of diversity can be accomplished by error prone PCR from a starting gene or the whole library can be assembled from gene fragments or from oligonucleotides, with the possibility to control completely the type and the number of mutations. Libraries of this kind are often created nowadays to address specific problems when structural information is available. For directed evolution of stability, or for new functions, or when a structure is not available, random mutagenesis is still preferred to produce more diversified variants.

Part IV: Methods of directed evolution

Directed evolution can be apparently considered as an antithesis to knowledge-based protein design approaches. In its general form, it relies on generation of diversity and a selection system to obtain new molecules with the desired properties. The procedure is based on cycles of selection and amplification of the selected members, creating a new enriched pool as start for the next round. The crucial point is the link between genotype and phenotype: it has to be ensured that only the coding sequences of the selected protein will go to the next cycles. If the diversity is sufficient and the selection system suitable, molecules with the

desired properties emerge from the selection, often with features not even predictable using rational approaches. The method reproduces rapidly an evolution process, hence the name.

The initial pool can be a designed library or may be generated by error prone PCR [71], DNA shuffling [72], guided recombination (e.g. the recent SCHEMA approach by Meyer *et al.* [73]) or similar methods. Additional diversity may be introduced at each cycle during the amplification of the recovered genetic material.

Selection systems can be divided in *in vitro* (e.g. ribosome display [26], mRNA display [74], cis display [75]), where no living cell is involved, and *in vivo* (e.g. phage display [76], bacterial display [77], yeast display [78]). In both cases the interaction at the base of the selection takes place *in vitro*, but in the *in vivo* systems the genetic material is maintained and propagated inside living cells; only the selected candidates transfer the genetic material to the next generation. In the *in vitro* systems the genetic material is accessible and new diversity can be directly generated during the amplification of the selected molecules. In the *in vivo* systems, additional steps are required to introduce new diversity. The genetic material needs first to be extracted and then reintroduced in the living system after increase of diversity. *In vivo* systems are generally more robust, but the library size is limited by the efficiency of transformation or transfection.

Apparently directed evolution can easily select molecules with the desired properties, finding sequences not even predictable by protein engineers. However, certain proteins might be incompatible with particular selection systems or maybe there is no suitable selection system for the properties that the proteins should be selected for. In these cases a design approach can represent an alternative way and also a starting point for selection.

Excluding these exceptions, when using directed evolution methods, the outcome is dependent not only on the selection, but also on the quality of the initial library. Library members able to unspecifically interact with some components of the selection system or with target molecules will thus be always selected, hiding the real desired molecules. At the same time, the more misfolded proteins are encoded in the library, the higher is the chance of unspecific binding, due to the exposed hydrophobic residues. Library members containing mutations in few precise positions not affecting the core packing are likely to be folded, but mutants with the whole sequence changed are more prone to incorrect folding, aggregation and unspecific binding. The latter can be indeed the case when diversity is randomly generated, for example using error prone PCR with high mutation rate. If a library is made from a stable scaffold, choosing which positions to mutate and avoiding mutations in the

hydrophobic core, some combination of residues can be missed but the chance of obtaining misfolded proteins is reduced and the overall library quality is higher.

Therefore, rational design and direct evolution are not necessarily in contrast, but can be combined together. Directed evolution on tailor made libraries, based on an engineered scaffold, can represent the best approach to obtain a molecule with the desired properties. Following such principle, a library was designed in this work based on an engineered armadillo repeat protein scaffold (Chapter 2); the procedure of library generation is described in Chapter 3. Use of ribosome display as selection method will be described in Chapter 4.

References

1. Pawson, T. & Nash, P. (2003). Assembly of cell regulatory systems through protein interaction domains. *Science* **300**, 445-52.
2. Pawson, T. & Scott, J. D. (1997). Signaling through scaffold, anchoring, and adaptor proteins. *Science* **278**, 2075-80.
3. Kuriyan, J. & Cowburn, D. (1997). Modular peptide recognition domains in eukaryotic signaling. *Annu Rev Biophys Biomol Struct* **26**, 259-88.
4. Letunic, I., Copley, R. R., Pils, B., Pinkert, S., Schultz, J. & Bork, P. (2006). SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res* **34**, D257-60.
5. Schultz, J., Milpetz, F., Bork, P. & Ponting, C. P. (1998). SMART, a simple modular architecture research tool: identification of signaling domains. *Proc Natl Acad Sci U S A* **95**, 5857-64.
6. Yaffe, M. B. (2002). Phosphotyrosine-binding domains in signal transduction. *Nat Rev Mol Cell Biol* **3**, 177-86.
7. Schlessinger, J. & Lemmon, M. A. (2003). SH2 and PTB domains in tyrosine kinase signaling. *Sci STKE* **2003**, RE12.
8. Waksman, G., Shoelson, S. E., Pant, N., Cowburn, D. & Kuriyan, J. (1993). Binding of a high affinity phosphotyrosyl peptide to the Src SH2 domain: crystal structures of the complexed and peptide-free forms. *Cell* **72**, 779-90.
9. Zhou, M. M., Huang, B., Olejniczak, E. T., Meadows, R. P., Shuker, S. B., Miyazaki, M., Trub, T., Shoelson, S. E. & Fesik, S. W. (1996). Structural basis for IL-4 receptor phosphopeptide recognition by the IRS-1 PTB domain. *Nat Struct Biol* **3**, 388-93.
10. Doyle, D. A., Lee, A., Lewis, J., Kim, E., Sheng, M. & MacKinnon, R. (1996). Crystal structures of a complexed and peptide-free membrane protein-binding domain: molecular basis of peptide recognition by PDZ. *Cell* **85**, 1067-76.
11. Yaffe, M. B., Rittinger, K., Volinia, S., Caron, P. R., Aitken, A., Leffers, H., Gamblin, S. J., Smerdon, S. J. & Cantley, L. C. (1997). The structural basis for 14-3-3:phosphopeptide binding specificity. *Cell* **91**, 961-71.
12. Lim, W. A., Richards, F. M. & Fox, R. O. (1994). Structural determinants of peptide-binding orientation and of sequence specificity in SH3 domains. *Nature* **372**, 375-9.
13. Huang, X., Poy, F., Zhang, R., Joachimiak, A., Sudol, M. & Eck, M. J. (2000). Structure of a WW domain containing fragment of dystrophin in complex with beta-

dystroglycan. *Nat Struct Biol* **7**, 634-8.
14. Nourry, C., Grant, S. G. & Borg, J. P. (2003). PDZ domain proteins: plug and play! *Sci STKE* **2003**, RE7.
15. Aitken, A. (2006). 14-3-3 proteins: a historic overview. *Semin Cancer Biol* **16**, 162-72.
16. Zarrinpar, A., Bhattacharyya, R. P. & Lim, W. A. (2003). The structure and function of proline recognition domains. *Sci STKE* **2003**, RE8.
17. Rudolph, M. G., Stanfield, R. L. & Wilson, I. A. (2006). How TCRs bind MHCs, peptides, and coreceptors. *Annu Rev Immunol* **24**, 419-66.
18. McFarland, B. J. & Beeson, C. (2002). Binding interactions between peptides and proteins of the class II major histocompatibility complex. *Med Res Rev* **22**, 168-203.
19. Rudolph, M. G., Speir, J. A., Brunmark, A., Mattsson, N., Jackson, M. R., Peterson, P. A., Teyton, L. & Wilson, I. A. (2001). The crystal structures of K(bm1) and K(bm8) reveal that subtle changes in the peptide environment impact thermostability and alloreactivity. *Immunity* **14**, 231-42.
20. Corper, A. L., Stratmann, T., Apostolopoulos, V., Scott, C. A., Garcia, K. C., Kang, A. S., Wilson, I. A. & Teyton, L. (2000). A structural framework for deciphering the link between I-Ag7 and autoimmune diabetes. *Science* **288**, 505-11.
21. Kajava, A. V. (2001). Review: proteins with repeated sequence--structural prediction and modeling. *J Struct Biol* **134**, 132-44.
22. Marcotte, E. M., Pellegrini, M., Yeates, T. O. & Eisenberg, D. (1999). A census of protein repeats. *J Mol Biol* **293**, 151-60.
23. Kobe, B. & Kajava, A. V. (2000). When protein folding is simplified to protein coiling: the continuum of solenoid protein structures. *Trends Biochem Sci* **25**, 509-15.
24. Groves, M. R. & Barford, D. (1999). Topological characteristics of helical repeat proteins. *Curr Opin Struct Biol* **9**, 383-9.
25. Andrade, M. A., Petosa, C., O'Donoghue, S. I., Müller, C. W. & Bork, P. (2001). Comparison of ARM and HEAT protein repeats. *J Mol Biol* **309**, 1-18.
26. Andrade, M. A., Perez-Iratxeta, C. & Ponting, C. P. (2001). Protein repeats: structures, functions, and evolution. *J Struct Biol* **134**, 117-31.
27. Gettemans, J., Meerschaert, K., Vandekerckhove, J. & De Corte, V. (2003). A kelch beta propeller featuring as a G beta structural mimic: reinventing the wheel? *Sci STKE* **2003**, PE27.
28. Hatzfeld, M. (1999). The armadillo family of structural proteins. *Int Rev Cytol* **186**, 179-224.
29. Blatch, G. L. & Lässle, M. (1999). The tetratricopeptide repeat: a structural motif mediating protein-protein interactions. *Bioessays* **21**, 932-9.
30. D'Andrea, L. D. & Regan, L. (2003). TPR proteins: the versatile helix. *Trends Biochem Sci* **28**, 655-62.
31. Smith, T. F., Gaitatzes, C., Saxena, K. & Neer, E. J. (1999). The WD repeat: a common architecture for diverse functions. *Trends Biochem Sci* **24**, 181-5.
32. Adams, J., Kelso, R. & Cooley, L. (2000). The kelch repeat superfamily of proteins: propellers of cell function. *Trends Cell Biol* **10**, 17-24.
33. Fontes, M. R., Teh, T. & Kobe, B. (2000). Structural basis of recognition of monopartite and bipartite nuclear localization sequences by mammalian importin-alpha. *J Mol Biol* **297**, 1183-94.
34. Scheufler, C., Brinker, A., Bourenkov, G., Pegoraro, S., Moroder, L., Bartunik, H., Hartl, F. U. & Moarefi, I. (2000). Structure of TPR domain-peptide complexes: critical elements in the assembly of the Hsp70-Hsp90 multichaperone machine. *Cell* **101**, 199-210.

35. Orlicky, S., Tang, X., Willems, A., Tyers, M. & Sicheri, F. (2003). Structural basis for phosphodependent substrate selection and orientation by the SCFCdc4 ubiquitin ligase. *Cell* **112**, 243-56.
36. Ho, M. S., Tsai, P. I. & Chien, C. T. (2006). F-box proteins: the key to protein degradation. *J Biomed Sci* **13**, 181-91.
37. Pancer, Z., Amemiya, C. T., Ehrhardt, G. R., Ceitlin, J., Gartland, G. L. & Cooper, M. D. (2004). Somatic diversification of variable lymphocyte receptors in the agnathan sea lamprey. *Nature* **430**, 174-80.
38. Coates, J. C. (2003). Armadillo repeat proteins: beyond the animal kingdom. *Trends Cell Biol* **13**, 463-71.
39. Lechner, E., Achard, P., Vansiri, A., Potuschak, T. & Genschik, P. (2006). F-box proteins everywhere. *Curr Opin Plant Biol* **9**, 631-8.
40. Dievart, A. & Clark, S. E. (2004). LRR-containing receptors regulating plant development and defense. *Development* **131**, 251-61.
41. Bork, P., Holm, L. & Sander, C. (1994). The immunoglobulin fold. Structural classification, sequence patterns and common core. *J Mol Biol* **242**, 309-20.
42. Almagro, J. C. (2004). Identification of differences in the specificity-determining residues of antibodies that recognize antigens of different size: implications for the rational design of antibody repertoires. *J Mol Recognit* **17**, 132-43.
43. MacCallum, R. M., Martin, A. C. & Thornton, J. M. (1996). Antibody-antigen interactions: contact analysis and binding site topography. *J Mol Biol* **262**, 732-45.
44. Marchalonis, J. J., Adelman, M. K., Robey, I. F., Schluter, S. F. & Edmundson, A. B. (2001). Exquisite specificity and peptide epitope recognition promiscuity, properties shared by antibodies from sharks to humans. *J Mol Recognit* **14**, 110-21.
45. Wilson, I. A., Ghiara, J. B. & Stanfield, R. L. (1994). Structure of anti-peptide antibody complexes. *Res Immunol* **145**, 73-8.
46. Hoogenboom, H. R. (2005). Selecting and screening recombinant antibody libraries. *Nat Biotechnol* **23**, 1105-16.
47. Harris, L. J., Larson, S. B., Hasel, K. W. & McPherson, A. (1997). Refined structure of an intact IgG2a monoclonal antibody. *Biochemistry* **36**, 1581-97.
48. Rini, J. M., Schulze-Gahmen, U. & Wilson, I. A. (1992). Structural evidence for induced fit as a mechanism for antibody-antigen recognition. *Science* **255**, 959-65.
49. Binz, H. K., Amstutz, P. & Plückthun, A. (2005). Engineering novel binding proteins from nonimmunoglobulin domains. *Nat Biotechnol* **23**, 1257-68.
50. Skerra, A. (2007). Alternative non-antibody scaffolds for molecular recognition. *Curr Opin Biotechnol* **18**, 295-304.
51. Bloom, J. D., Labthavikul, S. T., Otey, C. R. & Arnold, F. H. (2006). Protein stability promotes evolvability. *Proc Natl Acad Sci U S A* **103**, 5869-74.
52. Schimmele, B. & Plückthun, A. (2005). Engineering proteins for stability and efficient folding. In *Protein folding handbook* (Buchner, J. & Kiefhaber, T., eds.), pp. 1251-1303. Wiley-VCH, Weinheim.
53. Lehmann, M., Pasamontes, L., Lassen, S. F. & Wyss, M. (2000). The consensus concept for thermostability engineering of proteins. *Biochim Biophys Acta* **1543**, 408-415.
54. Socolich, M., Lockless, S. W., Russ, W. P., Lee, H., Gardner, K. H. & Ranganathan, R. (2005). Evolutionary information for specifying a protein fold. *Nature* **437**, 512-8.
55. Voigt, C. A., Gordon, D. B. & Mayo, S. L. (2000). Trading accuracy for speed: A quantitative comparison of search algorithms in protein sequence design. *J Mol Biol* **299**, 789-803.

56. Butterfoss, G. L. & Kuhlman, B. (2006). Computer-based design of novel protein structures. *Annu Rev Biophys Biomol Struct* **35**, 49-65.
57. Dahiyat, B. I. & Mayo, S. L. (1997). Probing the role of packing specificity in protein design. *Proc Natl Acad Sci U S A* **94**, 10172-7.
58. Dahiyat, B. I. & Mayo, S. L. (1997). De novo protein design: fully automated sequence selection. *Science* **278**, 82-7.
59. Desjarlais, J. R. & Handel, T. M. (1995). De novo design of the hydrophobic cores of proteins. *Protein Sci* **4**, 2006-18.
60. Lazar, G. A., Desjarlais, J. R. & Handel, T. M. (1997). De novo design of the hydrophobic core of ubiquitin. *Protein Sci* **6**, 1167-78.
61. Kono, H., Nishiyama, M., Tanokura, M. & Doi, J. (1998). Designing the hydrophobic core of Thermus flavus malate dehydrogenase based on side-chain packing. *Protein Eng* **11**, 47-52.
62. Mooers, B. H., Datta, D., Baase, W. A., Zollars, E. S., Mayo, S. L. & Matthews, B. W. (2003). Repacking the Core of T4 lysozyme by automated design. *J Mol Biol* **332**, 741-56.
63. Dunbrack, R. L., Jr. (2002). Rotamer libraries in the 21st century. *Curr Opin Struct Biol* **12**, 431-40.
64. Offredi, F., Dubail, F., Kischel, P., Sarinski, K., Stern, A. S., Van de Weerdt, C., Hoch, J. C., Prosperi, C., Francois, J. M., Mayo, S. L. & Martial, J. A. (2003). De novo backbone and sequence design of an idealized alpha/beta-barrel protein: evidence of stable tertiary structure. *J Mol Biol* **325**, 163-74.
65. Forrer, P., Stumpp, M. T., Binz, H. K. & Plückthun, A. (2003). A novel strategy to design binding molecules harnessing the modular nature of repeat proteins. *FEBS Lett* **539**, 2-6.
66. Stumpp, M. T., Forrer, P., Binz, H. K. & Plückthun, A. (2003). Designing repeat proteins: modular leucine-rich repeat protein libraries based on the mammalian ribonuclease inhibitor family. *J Mol Biol* **332**, 471-87.
67. Main, E. R., Xiong, Y., Cocco, M. J., D'Andrea, L. & Regan, L. (2003). Design of stable alpha-helical arrays from an idealized TPR motif. *Structure (Camb)* **11**, 497-508.
68. Nikkhah, M., Jawad-Alami, Z., Demydchuk, M., Ribbons, D. & Paoli, M. (2006). Engineering of beta-propeller protein scaffolds by multiple gene duplication and fusion of an idealized WD repeat. *Biomol Eng* **23**, 185-94.
69. Binz, H. K., Amstutz, P., Kohl, A., Stumpp, M. T., Briand, C., Forrer, P., Grütter, M. G. & Plückthun, A. (2004). High-affinity binders selected from designed ankyrin repeat protein libraries. *Nat Biotechnol* **22**, 575-82.
70. Rohl, C. A., Strauss, C. E., Misura, K. M. & Baker, D. (2004). Protein structure prediction using Rosetta. *Methods Enzymol* **383**, 66-93.
71. Fromant, M., Blanquet, S. & Plateau, P. (1995). Direct random mutagenesis of gene-sized DNA fragments using polymerase chain reaction. *Anal Biochem* **224**, 347-53.
72. Stemmer, W. P. (1994). DNA shuffling by random fragmentation and reassembly: in vitro recombination for molecular evolution. *Proc Natl Acad Sci U S A* **91**, 10747-51.
73. Meyer, M. M., Hochrein, L. & Arnold, F. H. (2006). Structure-guided SCHEMA recombination of distantly related beta-lactamases. *Protein Eng Des Sel* **19**, 563-70.
74. Roberts, R. W. & Szostak, J. W. (1997). RNA-peptide fusions for the in vitro selection of peptides and proteins. *Proc Natl Acad Sci U S A* **94**, 12297-302.
75. Odegrip, R., Coomber, D., Eldridge, B., Hederer, R., Kuhlman, P. A., Ullman, C., FitzGerald, K. & McGregor, D. (2004). CIS display: In vitro selection of peptides from libraries of protein-DNA complexes. *Proc Natl Acad Sci U S A* **101**, 2806-10.

76. Russel, M., Lowman, H. B. & Clackson, T. (2004). Introduction to phage biology and phage display. In *Phage display : a practical approach* (Clackson, T. & Lowman, H. B., eds.), Vol. 266, pp. xxiv, 332 p. Oxford University Press, Oxford.
77. Samuelson, P., Gunneriusson, E., Nygren, P. A. & Stahl, S. (2002). Display of proteins on bacteria. *J Biotechnol* **96**, 129-54.
78. Boder, E. T. & Wittrup, K. D. (1997). Yeast surface display for screening combinatorial polypeptide libraries. *Nat Biotechnol* **15**, 553-7.

Chapter 2
Natural and Designed Armadillo Repeat Proteins

What is an armadillo?

The most frequent question concerning armadillo repeat proteins is if they are actually coming from the animal armadillo. A simple answer is that they have no relationship with the animal, even though armadillos, like all the eukaryotic species, possess armadillo repeat proteins.

Historically the name armadillo was attributed to a *Drosophila melanogaster* mutant with a defect in the larval development. The mutant was identified during a series of pioneering experiments of systematic mutagenesis by Christiane Nüsslein-Volhard and Eric Wieschaus in the eighties, aimed to discover genes involved in embryogenesis [1; 2]. Their work was acknowledged with the Nobel Prize in 1995.

The body of the *D. melanogaster* larva is characterized by a precise segmentation pattern along the anterior-posterior axis (Fig. 1). Mutations in the armadillo gene affect every segment in the larva: the phenotype observed in the mutant is a duplication of the anterior part of each segment replaced the posterior part, forming a continuous series of stripes on the cuticle of the larva (Fig. 2). The phenotype classifies it as a "segment polarity gene" and accounts also for the name.

Fig. 1 Armadillo expression pattern in *Drosophila melanogaster* embryo. The green fluorescence indicates the presence of the armadillo protein. Image from Tolwinski group, Sloan-Kettering Institute, New York, USA.

In the following years the DNA and protein sequence analysis of armadillo revealed the presence of a new type of repeated sequence, hence named armadillo repeat [3]. New proteins were later discovered to have similar repeated sequences and named together as armadillo repeat proteins; the first definition of a common consensus sequence among the different proteins was given by Peifer *et al.* in 1994 [4].

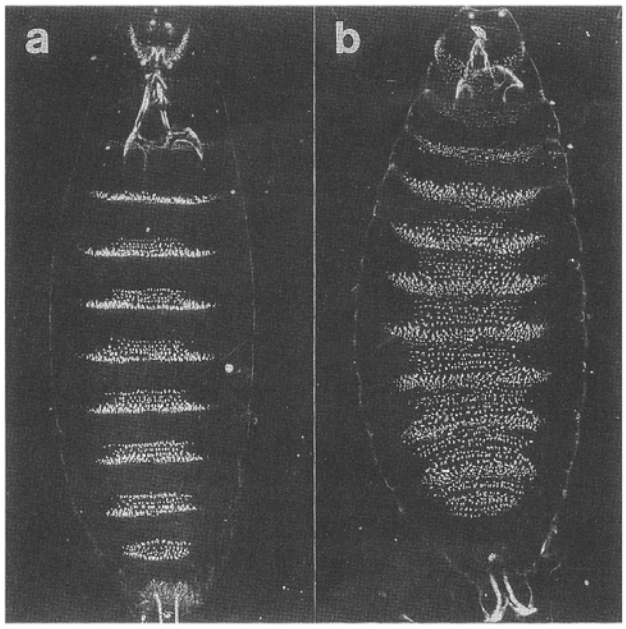


Fig. 2 *arm* embryos exhibit a segment polarity phenotype. Dark-field photomicrographs of ventral cuticle preparations of a wild-type embryo (a) and an *arm 2sB* embryo (b). Loss of arm function produces anterior-type denticles in the posterior naked cuticle region of each segment, often forming an apparent mirror image of the anterior denticle belt. Figure and description from Riggleman *et al.* [3].

The discovery of new proteins containing armadillo repeats increased further, but their three-dimensional structure remained elusive until 1997, when the structure of β-catenin, the mammalian homologue of armadillo, was unraveled by the group of William Weis [5], and shortly after that Conti *et al.* released the structure of the yeast karyopherin α [6].

Based on the available sequence information and on the structures, the armadillo repeat protein family was then more clearly defined in 2001 by Andrade *et al.* [7]. Their characterization of key residues and their definition of subgroups inside the family were the starting point for the design of an artificial armadillo repeat protein.

The armadillo repeat protein family

Armadillo repeat proteins form a widespread eukaryotic protein family found in every eukaryotic organism. Pfam [8] and SMART [9; 10] report approximately 1130 sequences containing armadillo repeat proteins.

Structurally and evolutionarily, armadillo repeat proteins are related to HEAT repeat proteins and a common origin has been proposed [11; 12]. However a detailed analysis of key residues for the three-dimensional structure allows one to distinguish between the two families and helps to identify subfamilies [7]. Armadillo repeat proteins are involved in protein-protein recognition in cytoplasm-nucleus transport (e.g. importin α), cytoskeleton regulation and anchoring to membrane structures (p120 family), signal transduction (β-catenin) and a

plethora of other function, especially characteristic of plants, where armadillo repeat proteins form also the recognition domain of several ubiquitin ligase complexes [13].

Nowadays 32 crystal structures are available but most of them are complexes of few proteins with different targets: β-catenin from mouse and human are almost identical (1 amino acid difference) and are present with 7 structures each, together with karyopherin α from *S. cerevisiae* (7 structures), importin α2 from mouse (9 structures), importin α2 from human (1 structure), plakophilin from human (1 structure). Other similar structures (e.g. pumilio repeat protein [14], Mo25α [15], mDia1 [16]) have been solved but their reduced sequence homology did not lead to an inclusion in the armadillo family. Solenoid structures composed of α-helices are recurring motives during evolution; whether they are a product of convergent evolution or quick divergence is still an open question.

S. cerevisiae contains only two armadillo repeat proteins, importin α and vac8, involved in membrane fusion and vacuole formation. Surprisingly, in particular for importin α, the number of repeats is the same as in plants, worms, arthropods or mammals, indicating that probably the armadillo family did not evolve from this stage by duplication of sequences. This hypothesis is supported when looking at the homology of the single repeats: each of them is more similar to the corresponding repeat in the orthologous proteins than to other repeats in the same protein, a clear indication of evolution of the domain as a single entity. The same observation holds for the other subfamilies. The presence of exon-intron boundaries not in register with the repeats confirms that the domain formation probably happened before the appearance of introns, and no further repeat expansion took place since then (Fig. 3).

How did this family arise in the first place in eukaryotes is still unknown, only few hypothetical proteins containing dispersed armadillo repeat related sequences can be found in archea and bacteria.

A coevolution of importin α with importin β (a HEAT repeat protein) can be postulated based on function and sequence similarity [11], but the evolutionary relationship between importin α, β-catenin and other armadillo repeat proteins, ubiquitous in multicellular organisms, is still obscure. A phylogenetic tree based on sequence distance (Fig. 4) reveals that the subfamily connections can be only traced back to a hypothetical ancestor of all the proteins containing armadillo repeats. Probably, the subfamilies evolved independently from each other.

Despite the scarce knowledge of armadillo repeat protein evolution, the sequence information represents the starting point for a consensus-based approach for the design of new

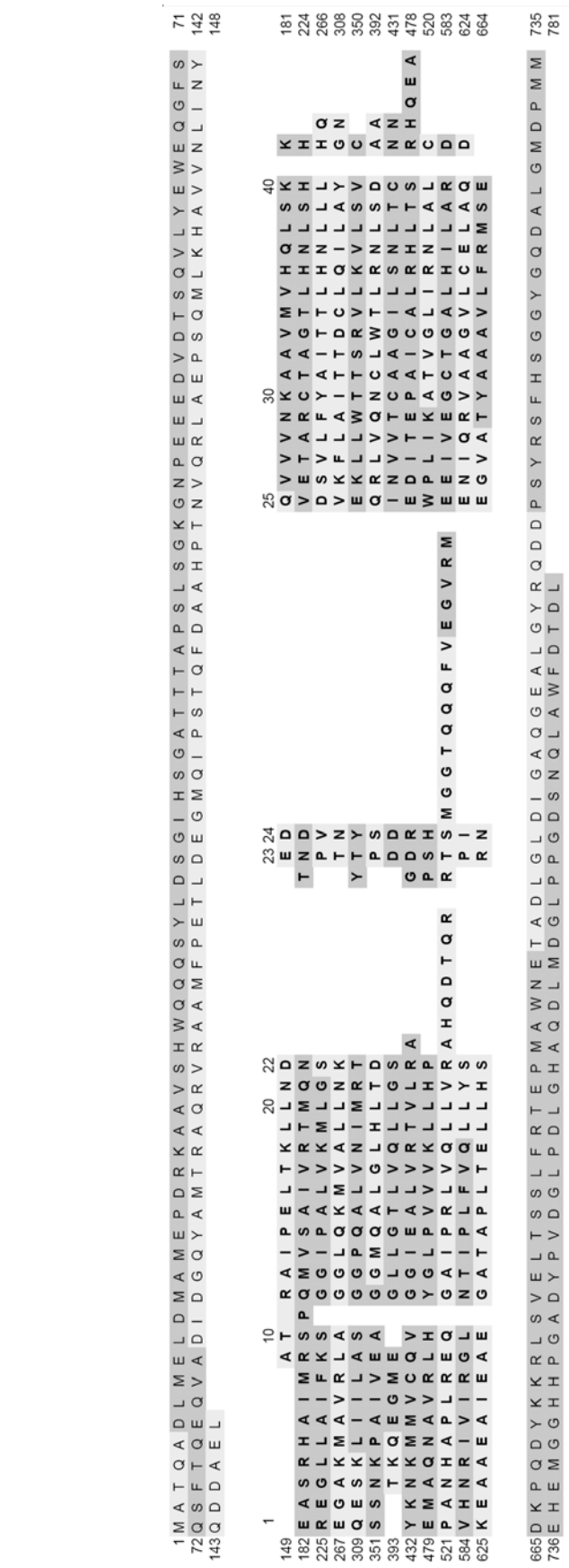


Fig. 3 Exon boundaries in the murine β-catenin. On the protein sequence, the regions encoded by each exon are indicated by alternating gray intensities. The armadillo repeats are aligned according to the structural information and the N- and C-terminal (residues 1-148 and 665-736, respectively) of the proteins are shown separately. Intron-exon boundaries are not conserved across repeats and no specific pattern can be detected to support the hypothesis of evolution by gene duplication.

armadillo repeat proteins. The sequences of natural armadillo domains do not easily allow the replacement, deletion, insertion of repeats because of lack of a common surface at the interfaces between repeats. Compatible repeats should then be generated, able to be combined and to grant improved biophysical properties to newly designed armadillo repeat proteins.

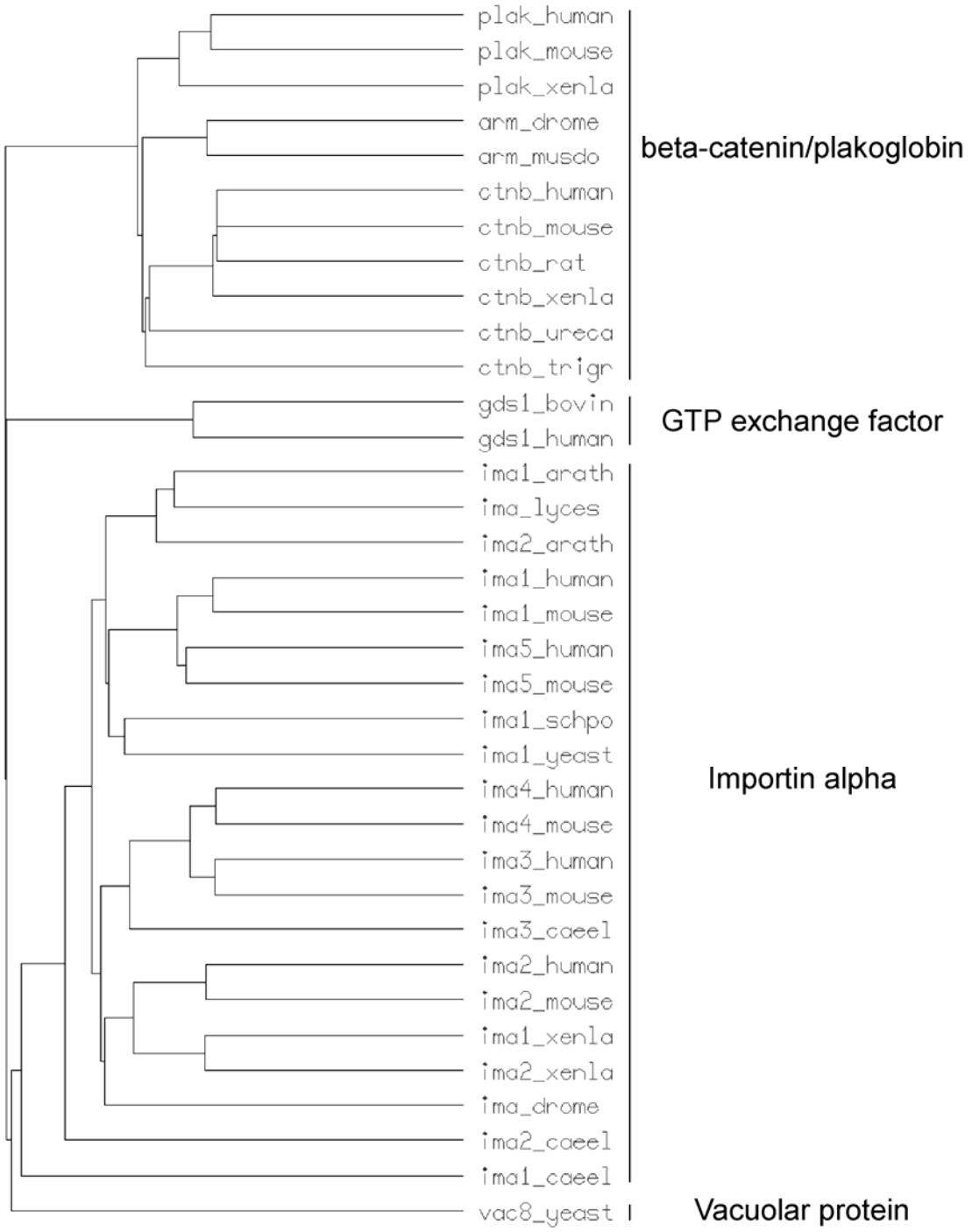


Fig. 4 Phylogenetic tree of armadillo repeat protein family. This three is a simplified version of the family tree, generated using representative sequences of different subfamilies. The proteins are identified by their UniProt entries. The phylogenetic tree was calculated from sequence homology using GCG (Accelrys Inc., San Diego, CA).

Designed armadillo repeat proteins as general peptide-binding scaffolds: consensus design and computational optimization of the hydrophobic core

Fabio Parmeggiani, Riccardo Pellarin, Anders Peter Larsen, Gautham Varadamsetty, Michael T. Stumpp, Oliver Zerbe, Amedeo Caflisch and Andreas Plückthun

Published in: Journal of Molecular Biology, **376**, 1282–1304 (2008)

doi:10.1016/j.jmb.2007.12.014 *J. Mol. Biol.* (2008) **376**, 1282–1304

JMB

Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Designed Armadillo Repeat Proteins as General Peptide-Binding Scaffolds: Consensus Design and Computational Optimization of the Hydrophobic Core

Fabio Parmeggiani[1], Riccardo Pellarin[1], Anders Peter Larsen[1], Gautham Varadamsetty[1], Michael T. Stumpp[1], Oliver Zerbe[2], Amedeo Caflisch[1] and Andreas Plückthun[1]*

[1]*Department of Biochemistry, University of Zürich, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland*

[2]*Department of Organic Chemistry, University of Zürich, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland*

Received 3 July 2007; received in revised form 13 November 2007; accepted 5 December 2007
Available online 14 December 2007

Edited by F. E. Cohen

Armadillo repeat proteins are abundant eukaryotic proteins involved in several cellular processes, including signaling, transport, and cytoskeletal regulation. They are characterized by an armadillo domain, composed of tandem armadillo repeats of approximately 42 amino acids, which mediates interactions with peptides or parts of proteins in extended conformation. The conserved binding mode of the peptide in extended form, observed for different targets, makes armadillo repeat proteins attractive candidates for the generation of modular peptide-binding scaffolds. Taking advantage of the large number of repeat sequences available, a consensus-based approach combined with a force field-based optimization of the hydrophobic core was used to derive soluble, highly expressed, stable, monomeric designed proteins with improved characteristics compared to natural armadillo proteins. These sequences constitute the starting point for the generation of designed armadillo repeat protein libraries for the selection of peptide binders, exploiting their modular structure and their conserved binding mode.

Keywords: consensus design; armadillo repeat; hydrophobic core; molten globule; molecular dynamics and minimization

Introduction

In recent years, as an alternative to raising monoclonal antibodies by immunization, recombinant antibodies[1] and an increasing number of other protein scaffolds[2] have been investigated as novel binding molecules. However, neither antibodies themselves nor any of these alternative protein scaffolds were specifically designed to bind peptides. Target-specific binding molecules are, in general, obtained from large protein libraries by *in vitro* selection or, in the case of monoclonal antibodies, through traditional immunization procedures. Both approaches require that, for each target, each new binding molecule is individually generated and characterized for specificity and

*Corresponding author. E-mail address: plueckthun@bioc.uzh.ch.

Present addresses: A.P. Larsen, Department of Biomedical Sciences, University of Copenhagen, Blegdamsvej 3, DK-2200 Copenhagen, Denmark; M.T. Stumpp, Molecular Partners AG, Grabenstrasse 11a, CH-8952 Schlieren, Switzerland.

Abbreviations used: αArm, Armadillo domain of human importin-α1 (residues 83–505); βArm, Armadillo domain of mouse β-catenin (residues 150–665); ANS, 1-anilino-naphthalene-8-sulfonate; C-type, overall consensus repeat; CD, circular dichroism; HA, hemagglutinin tag; HSQC, heteronuclear single quantum coherence; I-type, importin-derived consensus armadillo repeat; IMAC, immobilized metal-ion affinity chromatography; M-type, mutated armadillo repeat obtained by computational approach; MALS, multiangle light scattering; MRE, mean residue ellipticity; NLS, nuclear localization sequence; NOE, nuclear Overhauser enhancement; pD, phage lambda protein D; PDB, Protein Data Bank; SDS-PAGE, sodium dodecyl sulfate–polyacrylamide gel electrophoresis; SEC, size-exclusion chromatography; T-type, catenin/plakoglobin-derived consensus armadillo repeat.

0022-2836/$ - see front matter

cross-reactivity, making the generation of binders against a large number of peptide targets (e.g., representing a full proteome) an almost prohibitive task.

The aim of the present study was to develop a scaffold for the generation of peptide-specific binding proteins. In more detail, we wanted to develop proteins that were stable under various conditions and with the intrinsic ability to bind peptides in a conserved fashion. To recognize peptides in a sequence-selective manner the specificity of binding should ideally be conferred through specific interactions with the peptide side chains.

Natural peptide-binding scaffolds can be grouped in different classes. Antibodies are known to be able to bind peptides and have been well characterized.[3–6] Although peptide-binding antibodies have certain structural features in common, the mode of binding is not conserved. Thus, the information acquired through studies of antibody–peptide complexes cannot easily be applied to the general design of peptide-binding antibodies or extended to other proteins.

Small adaptor domains (e.g., SH2, SH3, and PDZ)[7] show specific binding to their targets, usually in a conserved binding mode within one family, but their affinity is generally low. The recognition sequence is very short and biased toward certain amino acid types, posttranslational modifications, or free N- or C-termini. While several such domains could be linked together by flexible peptides to recognize longer peptide sequences, a coverage of any arbitrary sequence would still be very difficult since these small domains might not be adaptable to the recognition of any arbitrary sequence. Furthermore, the entropy loss upon binding of such flexibly linked constructs would not necessarily lead to high affinities.

The major histocompatibility complex proteins (MHC I and MHC II)[8] possess a higher intrinsic variability and the ability to recognize a broad range of peptides, but the difficulties in their handling reduce their attractiveness as a scaffold candidate.

Repeat proteins, in particular tetratricopeptide repeats (TPRs),[9] armadillo,[10] and WD40[11] proteins, have been shown to possess an intrinsic ability to bind peptides, taking advantage of their repetitive structure. Thus, for our purpose, a scaffold based on repeat proteins seemed to constitute a promising candidate. For reasons outlined below, we chose the armadillo repeat protein family as the basis for our scaffold candidate.

Armadillo repeat proteins[12,13] are abundant in eukaryotes, where they are involved in a broad range of biological processes (e.g., transcription regulation,[14] cell adhesion,[15] tumor suppressor activity,[16] and nucleocytoplasmic transport[17]). These proteins are characterized by tandem repeats of approximately 42 amino acids that were first discovered in the product of the *Drosophila melanogaster* segmentation polarity gene Armadillo, which is homologous to mammalian β-catenin.[18,19] Armadillo repeat proteins participate in protein–protein interactions, and the armadillo domain is usually involved in the recognition process. The domain forms a right-handed superhelix[20,21] (Fig. 1a), as shown by the crystal structures of β-catenin[22] and importin-α.[23] Every repeat is composed of three α-helices, named H1, H2, and H3 (Fig. 1b), and several repeats stack to form the compact domain. Specialized repeats are present at the N- and C-termini of the protein, protecting the hydrophobic core from solvent exposure (Fig. 1a).

Armadillo repeat proteins are able to bind different types of peptides, yet relying on a conserved binding mode of the peptide backbone. Reported dissociation constant (K_d) values as low as 10–20 nM[24]

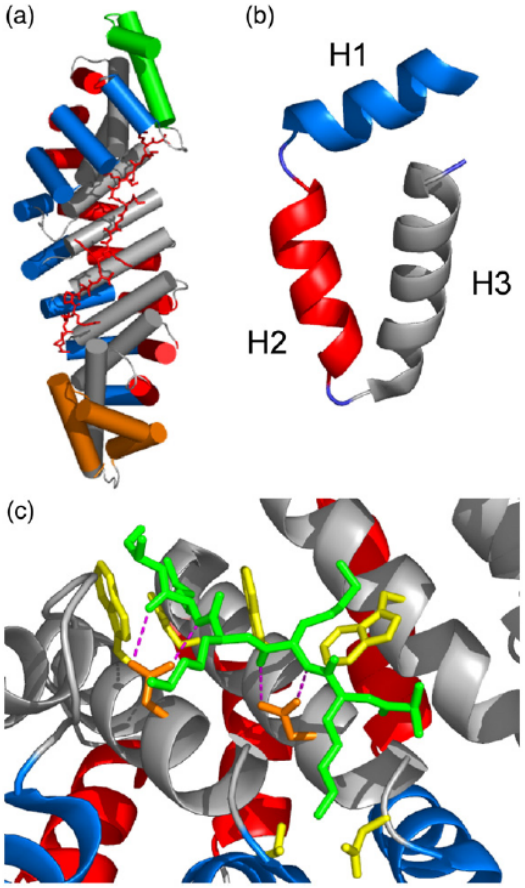


Fig. 1. (a) Structure of *S. cerevisiae* importin-α in complex with nucleoplasmin NLS (PDB ID 1EE5), showing the right-handed superhelical structure typical for armadillo repeat proteins. The cylinders represent the α-helices. The N-terminal repeat is indicated in green, and the C-terminal repeat is shown in orange. The bound peptide is depicted in red in a stick representation. (b) Detail of repeat 6 from 1EE5. The α-helices are represented as ribbons. (c) Detail of the peptide-binding mode. The conserved asparagine residues (in orange) contact *via* hydrogen bonds (purple) the backbone of the peptide, depicted in green. The residues that are responsible for the interactions with the side chains of the target peptide are shown in yellow. In all panels, helix 1 (H1) is indicated in blue, helix 2 (H2) in red, and helix 3 (H3) in gray.

indicate that high affinities can be achieved. Crystal structures of armadillo repeat proteins in complex with bound peptides have revealed that most peptide targets are bound in an extended conformation along the surface, inside the groove formed by the H3 helices. The superhelical armadillo domain winds around the peptide, oriented in the opposite N- to C-terminal direction (Fig. 1a), thus forming a double-helical complex, topologically similar to the DNA double strand. An asparagine residue, conserved in almost every repeat at the C-terminal part of H3, makes hydrogen bonds to the main chain of the target peptide, thereby keeping it in an extended conformation. Additional interactions to the target side chains are provided by neighboring residues, mostly in H3 (Fig. 1c). In a first approximation, each dipeptide unit of the target peptide is specifically recognized by one repeat in the armadillo domain (Fig. 2a).

In theory, the possibility of developing individual repeats that specifically bind a two-amino-acid sequence unit is very attractive. Given that the individual repeats are based on the same optimized scaffold and, thus, compatible with each other, any given number of repeats can be directly stacked to extend the recognition to much longer peptide sequences. In contrast to flexibly linked small adaptor domains mentioned above, armadillo repeats directly stack on each other in a rather rigid manner, allowing binding to uninterrupted longer peptides. This would exploit the specificity of the individual repeats to provide a peptide-binding designed armadillo protein with high and predetermined specificity, governed by the individual repeats. Such an approach (Fig. 2b), using armadillo proteins assembled from previously selected "building blocks," could effectively bypass the current *in vitro* selection procedures for individual peptides. However, this requires such individual peptide-specific repeats first to be developed, using a library-based approach.

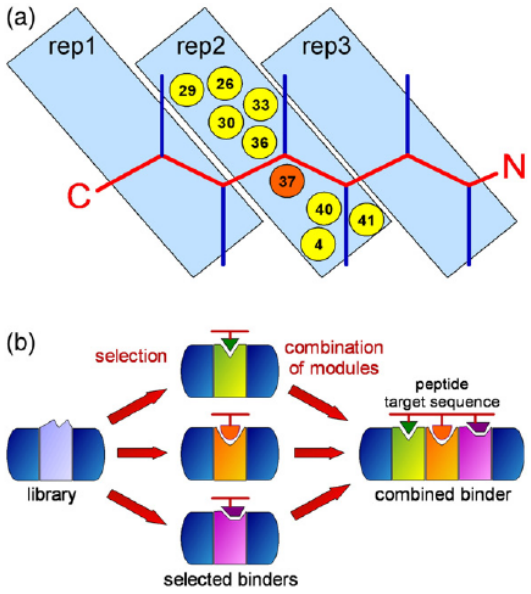


Fig. 2. Binding of target peptides. (a) Schematic drawing of an armadillo repeat protein binding to an extended peptide. The target peptide is bound in an antiparallel orientation to the protein. N and C indicate N- and C-termini of the peptide, which is depicted in red, with the amino acid side chains shown in blue. The residues of armadillo repeats involved in binding occupy specific positions within the single repeat sequences, mostly on helix 3. The position indicated in orange (a conserved Asn) is responsible for the binding of the peptide main chain; the positions in yellow are involved in recognition of the peptide side chains. (b) Designed armadillo repeat proteins potentially allow the selection of single repeats that specifically recognize short sequences. The selected peptide-specific repeats can be then combined to recognize longer peptides without performing additional selections.

In the present study, we have, as a first step, designed armadillo repeat modules based on consensus sequences. Proteins containing different types of modules have been assembled and characterized, initially only leading to stable dimeric proteins or monomeric molten-globule-like proteins. We subsequently used a combination of molecular dynamics and minimization to improve the hydrophobic core packing and convert the consensus-designed armadillo repeat protein with molten-globule-like properties to a monomeric, stable folded protein. Finally, the protein characteristics were evaluated for exploring the possibility of generating a modular peptide-binding scaffold. We succeeded in developing a stable, monomeric consensus protein that can be used now in the generation of peptide-specific individual armadillo repeat proteins.

Results

Armadillo repeat protein design

A consensus design strategy[25] has been applied in order to generate armadillo repeat proteins with high expression levels of soluble protein in *Escherichia coli*, monomeric state, high thermodynamic stability, and absence of cysteines for convenient expression and handling.

This design procedure was aimed at the generation of self-compatible repeat modules; therefore, consensus sequences were derived from multiple alignments of single armadillo repeats from the Swiss-Prot database.[26] A consensus design strategy has been successfully applied previously to other designed repeat proteins,[27–30] and it is based on consensus design of internal repeats (or internal modules). Special terminal capping repeats (terminal modules) have been generated to protect the hydrophobic core from solvent exposure. The crucial role of capping repeats has been previously shown in studies with designed ankyrin repeat proteins.[31]

The numbering used here to define the positions within the repeats was based on the family align-

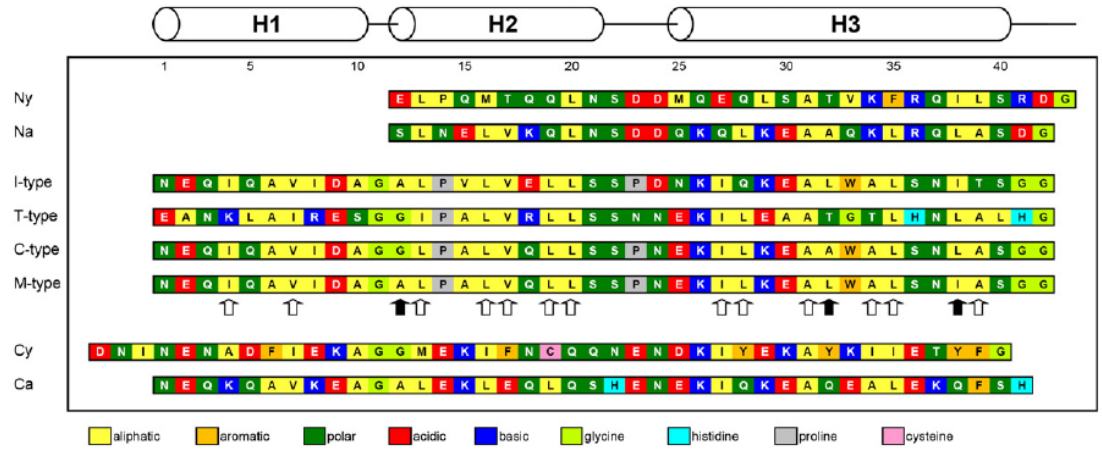


Fig. 3. Sequences of the designed internal modules and capping repeats; the cylinders indicate the helices, and the numbers denote the positions within the single repeats according to the convention introduced. *Ny* is an N-terminal capping repeat derived from importin-α from the *y*east *S. cerevisiae*; *Na* is an *a*rtificial N-terminal capping repeat. *I-type* is the internal module based on sequences from the *i*mportin-α subfamily, *T-type* is the internal module based on sequences from the β-cat*e*nin/plakoglobin subfamily, *C-type* is the overall *c*onsensus based on both subfamilies. *M-type* is the mutant sequence obtained through the computational approach described here. *Cy* is a C-terminal capping repeat derived from importin-α from the *y*east *S. cerevisiae*; *Ca* is an *a*rtificial C-terminal capping repeat. The amino acid color code is indicated below the sequences. The arrows indicate the positions considered in the computational approach. The filled arrows show the positions that differ between the C-type and the finally chosen M-type module.

ment proposed by Andrade *et al.*[32] Position −2 in that work corresponds to position 1 in the numbering used in the present study, where the putative helices H1, H2, and H3 encompass residues 1–10, 12–21, and 25–40, respectively.

Consensus design of internal modules

The initial sequence profile was generated using the family alignment from SMART†[33,34] (data from January 2004) as starting point (the consensus sequence is shown in Supplementary Fig. S1a). We largely followed the steps previously outlined.[27,30] We first removed all sequences lacking annotation in the Swiss-Prot database, especially hypothetical proteins or sequences for which no protein data were available with the exception of indirect evidence by sequence homology. The final set of 319 sequences led to a profile of 40 residues, covering the repeat sequence from H1 to H3 but excluding the loop between H3 and the next repeat. This sequence profile was used for a further search against the Swiss-Prot database. The repeats thus found belong to proteins that fall into different subfamilies of armadillo repeat proteins, as indicated by Andrade *et al.*[32]

Nevertheless, the sequences from different subfamilies might not be compatible. Taking this possibility into account, three final consensus sequences were constructed: one derived from β-catenin/plakoglobin (110 sequences), one from importin-α (133 sequences), and one from the combination of both (243 sequences). No normalization was applied during the calculation of the combined consensus sequence, which would compensate for a slight overrepresentation of importin-α over β-catenin/plakoglobin sequences in our selected set (133 over 110). The automatic alignments, performed with ClustalW[35] (Supplementary Fig. S1b), were manually refined including the loop connecting adjacent repeats (Supplementary Fig. S1c).

Structural information was taken into account to replace the cysteines present in the consensus sequences and reduce possible steric clashes. A more detailed description of additional sequence features and the rationale for amino acid exchanges are provided in the Supplementary Materials. Requirements for the cloning strategy were also considered at this stage, leading to the final module sequences type I (derived from *i*mportin-α subfamily), type T (derived from β-cat*e*nin/plakoglobin subfamily), and type C (*c*ombined consensus between these two subfamilies) (Fig. 3). Positions 7, 16, 17, 19, 20, 31, 34, 35, and 38 are well conserved in all the sequences and are part of the hydrophobic core of the armadillo proteins.

The positions potentially involved in binding of peptides (4, 26, 29, 30, 33, 36, 37, 40, and 41) have been defined based on the analysis of structures of complexes (summarized by Lange *et al.*[36] and Xu and Kimelman[37]) and data from mutation experiments.[38–40] The conserved Asn, responsible for binding to the main chain of the target peptide and at least in part for keeping it in an extended conformation, is located at position 37. Position 4 is part of both the hydrophobic core and the peptide binding site, and thus, the types of residues allowed at this position in a potential library would probably be restricted.

†http://smart.embl.de

Design of capping repeats

N- and C-terminal capping repeats, found in natural armadillo domains, protect the hydrophobic core, as they present a hydrophobic surface to the internal repeat side but a hydrophilic surface to the solvent. Capping sequences have also been considered in the previous design of other repeat proteins.[27,29–31]

The boundaries of armadillo domains have been estimated by limited proteolysis.[22,23] However, they are not clearly defined, partly due to the weak similarity of the terminal repeats to the internal ones. In addition, not all the residues are visible in the crystal structures of importin-α and β-catenin. It is likely that only the visible residues contribute to the armadillo domain, and the additional parts are unstructured and do not strictly belong to the domain. We have defined the N-terminal capping repeat as starting from position 12 (the beginning of H2). In contrast, the C-terminal capping repeat is completely resolved in the x-ray structures, and we defined it to comprise position 1 to position 41, thus including H1 to H3.

The capping repeats have been designed by using two different approaches. In the first, natural capping repeats were adapted to our designed internal repeats. Structural information to ensure compatibility between the capping repeats and the designed internal repeat is a fundamental prerequisite. The importin-α from *Saccharomyces cerevisiae* was considered to be the best candidate for a general capping repeat donor: all our designed modules present a flat surface that can interact with the inner surface of yeast importin-α capping repeats, as judged from molecular models. The *y*east importin-α-derived N-terminal and C-terminal capping repeats were named N*y* and C*y*, respectively.

The N-terminal capping repeat covers the residues from Glu88 to His119 of yeast importin-α. However, the two residues Glu118–His119 were replaced by Asp–Gly (Fig. 3, positions 42 and 43 of Ny) to adapt the terminal loop to the designed modules: glycine is used for assembly of the modules (as its codon overlaps a restriction site) and aspartate keeps a negative charge, which is frequently present at this position in natural proteins, reducing at the same time the helical propensity in the turn region.

The C-terminal capping repeat covers the region from Asn471 to Gly510 in yeast importin-α. However, the loop connecting the last internal repeat with the C-terminal repeat contains additional residues in yeast importin-α, compared to other natural importins. A modified version of this C-terminal capping repeat has thus been generated by introducing three residues (Asp-Asn-Ile) before H1 (Fig. 3, first three residues of Cy). Asn and Ile are naturally present at these positions; Asp has been included to keep a negatively charged loop as observed in several natural sequences while reducing the helical propensity.

In the second approach, two completely *a*rtificial N- and C-terminal capping repeats were designed (named N*a* and C*a*, respectively, and shown in Fig. 3), starting from the type C consensus and substituting the exposed hydrophobic residues with hydrophilic ones. Positions 12, 19, 27, and 34 of the N-terminal capping repeat are occupied by hydrophobic residues in the consensus sequence and were replaced by hydrophilic residues based on structures and frequently occurring residues obtained from alignment of N-terminal capping sequences. In a similar way, positions 8, 13, 17, 20, 28, 32, 35, 38, and 39 of the C-terminal capping repeat were replaced by hydrophilic residues based on structures and frequently occurring residues obtained from alignment of C-terminal capping sequences. A detailed description of the residues introduced in the designed capping repeats is provided in the Supplementary Materials. A second version of the Cy capping repeat was also designed without the three initial residues and with Ala replacing Cys at position 19; no change was observed, compared to Cy, in the level of expression and in the amount of soluble protein, and it will thus not be discussed further.

Assembly, cloning, and expression of designed armadillo repeat proteins

The amino acid sequences of all modules were back-translated to DNA sequences, optimizing the codon usage for expression in *E. coli*. Each module was synthesized, starting from overlapping oligonucleotides (Supplementary Table S1).

The modules were assembled stepwise using type IIS restriction enzymes (Supplementary Fig. S2), following the approach reported by Binz *et al.*[27] The final proteins were named according to the modules that they contain: the name indicates, in order, the type of N-terminal repeat (A for *A*rtificial or N*a*, Y for *Y*east derived or N*y*), the type of internal repeats (type I, type T, or type C) with the number of modules used as a subscript, and the type of C-terminal repeat (A for *A*rtificial or C*a*, Y for *Y*east derived or C*y*): for example, YI_4A contains a *Y*east-derived N-terminal repeat (N*y*), four internal repeats based on *I*mportin consensus (type *I*), and an *A*rtificial C-terminal repeat (C*a*).

Thus, Na or Ny as N-capping modules were combined with T-, I-, or C-type internal modules and Ca or Cy C-terminal modules, leading to 12 possible combinations. The proteins contain only one type of internal module to avoid incompatible surfaces at the interface between repeats. The influence of capping and internal repeats was evaluated by analyzing the expression properties of all the constructs, containing two or four internal repeats. The proteins were expressed in *E. coli* XL1-blue using a pQE30-based expression plasmid, providing an N-terminal $MRGSH_6$ tag for purification. The insert was constructed with a double stop codon (Supplementary Fig. S3). As an example, the DNA and protein sequences of YC_2A are provided in Supplementary Fig. S3.

The highest level of soluble protein expression was obtained when the internal modules were combined with Ny and Ca (Fig. 4a). The Na cap leads to almost undetectable expression in Coomassie-stained polyacrylamide gels, and the presence of Cy resulted in a substantial portion of the protein found in the insoluble fraction after cell lysis. The observed effects of terminal capping repeats were independent of the type and the number of internal modules. However, increasing the number of internal modules enhanced the amount of soluble protein and the absolute amount of protein produced. Remarkably, type T proteins are characterized by a lower apparent mobility in sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS-PAGE), compared to type I and type C proteins (Fig. 4b).

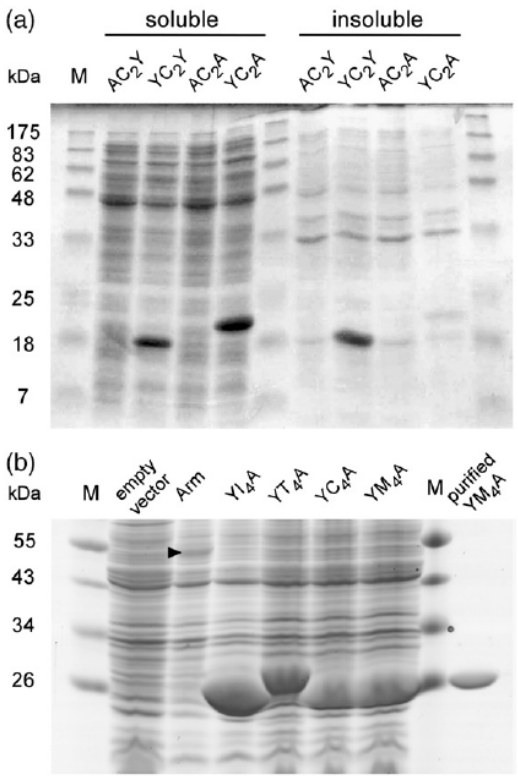


Fig. 4. (a) Influence of capping repeats on expression. Soluble and insoluble fractions of *E. coli* cell extracts are shown in a Coomassie-stained SDS polyacrylamide gel. The proteins contain two internal C-type modules with different combinations of capping repeats. (b) Whole-cell extracts of consensus proteins. The constructs contain Ny and Ca as capping repeats. Cells transformed with the empty vector or with the vector containing the armadillo domain of mouse β-catenin (Arm) were used as control. The proteins can be easily purified in a single step by IMAC, as shown for YM_4A. The expected size is 18 and 27 kDa for proteins containing two or four internal modules, respectively, and 56 kDa for Arm. The triangle indicates the band corresponding to the armadillo domain of β-catenin, which is expressed at much lower yield than the designed proteins. The molecular mass of the marker (M) is indicated in kilodaltons on the left.

Protein purification and characterization: Comparison with natural armadillo domains

Proteins containing the combination of Ny, Ca, and two, four, or eight internal repeats have been chosen for biophysical characterization and evaluation of the properties of type I, type T, and type C modules. The results are summarized in Table 1.

The purification by immobilized metal-ion affinity chromatography (IMAC) in a single step provided up to 100 mg of pure protein from 1 l of bacterial culture (Fig. 4b). No sign of precipitation or degradation was detected by spectrophotometry and SDS-PAGE in protein solutions stored for up to 1 month at 4 °C in the IMAC elution buffer.

The natural human importin-α1 (Swiss-Prot P52294) and the mouse β-catenin (Swiss-Prot Q02248) were also expressed using the same pQE30-based plasmid. The importin contains 10 armadillo repeats and the catenin 12, including the capping repeats in the count in both cases. Human importin-α1 gave the highest yield among the importin-α family members tested (data not shown) after a two-step coupled IMAC–ion exchange purification and, together with mouse β-catenin, was used for the comparison with the designed armadillo repeat proteins.

Importin-α1 and β-catenin, despite their elongated shape, elute at the volume expected from their molecular weight in gel filtration on a Superdex 200 column, and the monomeric state was confirmed for both proteins by multiangle light scattering (MALS) measurements (Table 1).

On the other hand, the designed proteins show elution volumes corresponding to higher-than-expected apparent molecular masses in size-exclusion chromatography (SEC) (Table 1). MALS indicates that the I- and T-type proteins are probably present as a mixture of dimers and monomers in solution. The main peak (Fig. 5a) corresponds to the dimeric form, and this value is reported in Table 1. At high concentration (2–4 mg/ml), I- and T-type proteins are present as a mixture of oligomers. In contrast, monomeric and oligomeric fractions of C-type proteins YC_4A (Fig. 5a) and YC_8A (data not shown) can be separated, up to the highest concentration tested (4 mg/ml). However, the fractions of YC_4A and YC_8A, shown by MALS to be monomeric, elute earlier than expected for proteins of comparable size. The smaller YC_2A represents the only exception: independent of the concentration, the MALS-calculated mass values are always intermediate between monomer and dimer. A decrease in pH to 7 favors the formation of oligomeric species of I- and T-type proteins. C-type proteins are, in contrast, unaffected by pH (data not shown).

The circular dichroism (CD) spectra (Fig. 5b) indicate the presence of significant α-helical secondary structure content for all proteins, particularly for the I-type proteins. For I- and C-type consensus repeats, the absolute value of mean residue ellipticity (MRE) and the helical content generally increase

Table 1. Biophysical properties of designed and natural armadillo repeat proteins

Construct	Residues (repeats)[a]	pI[b]	MW_{calc} (kDa)[b]	Oligomeric state[c]	MW_{obs} (kDa)[d]	$MW_{obs/calc}$[e]	CD_{222} (MRE)[f]	Helical content (%)[g]	Observed T_m (°C)[h]
YI_2A	169 (4)	5.2	18.6	Dimer	64.6	1.7	−13,000	63	~55
YI_4A	253 (6)	4.8	27.4	Dimer	116.1	2.1	−19,500	80	~69
YI_8A	421 (10)	4.6	44.9	Dimer	148.8	1.7	−22,600	85	>85
YT_2A	169 (4)	6.3	18.6	Dimer	141.2	3.8	−7100	23	~56
YT_4A	253 (6)	6.5	27.3	Dimer	219.6	4.0	−10,100	40	~75
YT_8A	421 (10)	6.7	44.8	Dimer	229.7	2.6	−9400	35	~83
YC_2A	169 (4)	5.4	18.4	Mixture	59.1	n.d.	−9100	45	n.d.
YC_4A	253 (6)	5.1	26.9	Monomer	50.0	1.9	−12,100	49	n.d.
YC_8A	421 (10)	4.8	44.0	Monomer	76.7	1.7	−20,000	62	n.d.
YM_4A	253 (6)	5.1	27.1	Monomer	32.2	1.2	−18,800	87	~70
αArm[i]	435 (10)	5.5	48.2	Monomer	42.9	0.9	−14,300	54	~43
βArm[j]	528 (12)	8.7	57.6	Monomer	52.6	0.9	−16,800	60	~58

n.d. indicates that the value has not been determined due to either an inhomogeneous sample (oligomeric state of YC_2A) or lack of cooperative transition in thermal denaturation (YC_2A, YC_4A, and YC_8A).

[a] The number of residues includes the $MRGSH_6$ tag; the number of repeats includes capping repeats.

[b] *pI* and molecular weight calculated from the sequence; masses were confirmed by mass spectrometry.

[c] Oligomeric state as indicated by multiangle static light scattering.

[d] Observed molecular weight as determined in SEC.

[e] Ratio between observed and calculated molecular weight, taking into account the oligomeric state (Os): $MW_{obs/calc} = MW_{obs}/(Os \cdot MW_{calc})$.

[f] Mean residue ellipticity at 222 nm expressed as $deg \cdot cm^2/dmol$.

[g] Helical content estimated with the program CDpro.[41]

[h] T_m observed in thermal denaturation by CD.

[i] Armadillo domain of human importin-α1.

[j] Armadillo domain of mouse β-catenin.

with the number of internal repeats; in contrast, the helical content is almost constant for T-type proteins (Supplementary Fig. S4). The values of helical content were calculated using the program CDpro[41] and are indicated in Table 1.

The CD signal at 222 nm was chosen to monitor stability against thermal and denaturant-induced unfolding. I- and T-type proteins show a cooperative transition, while no transition was observed in C-type proteins (Fig. 5c). The midpoint of transition during thermal denaturation (T_m) increases with the number of repeats, for example, from approximately 70 °C for YI_4A to more than 80 °C for YI_8A (Table 1). Importin-α1 and β-catenin, containing 8 and 10

Fig. 5. Biophysical characterization of designed and natural armadillo repeat proteins. (a) SEC and MALS of designed armadillo repeat proteins containing four internal modules and of importin-α1. YI_4A, YT_4A, and YC_4A show apparent molecular weights higher than the globular proteins with the same calculated mass (about 27 kDa). The broad peaks shown by YI_4A and YT_4A are due to a mixture of dimers and monomers, as indicated by the molecular mass determined by light scattering. The highest point of the peak corresponds to the dimeric fraction. In the case of YC_4A, the first peak eluted contains probably a mixture of oligomers with high molecular masses. The monomeric peak after separation remains monomeric and was further characterized. The importin-α1 (αArm) is a monomer as indicated by LS and elutes at the expected volume. The data were obtained with a Superdex 200 column. The elution was followed by absorbance at 280 nm for YC_4A, YI_4A and αArm; YT_4A does not possess any residue absorbing significantly at 280 nm; thus, the elution was followed at 230 nm. V_0 indicates the void volume of the column. Alcohol dehydrogenase (ADH; MW = 150 kDa), bovine serum albumin (BSA; MW = 66 kDa), carbonic anhydrase (CA; MW = 29 kDa), and aprotinin (Apr; MW = 6.5 kDa) were used as molecular weight markers, and the corresponding elution volumes are indicated by the arrows. (b) CD spectra of I-type, T-type, and C-type proteins containing four internal modules. The natural armadillo domains of human importin-α1 (αArm) and mouse β-catenin (βArm) are indicated by open and filled circles, respectively. The values are reported as MRE. (c) Thermal denaturation curves. A comparison between designed armadillo repeat proteins containing four or eight internal modules is shown, from the top, for I-type, T-type, and C-type proteins. αArm and βArm are displayed in the bottom panel. The denaturation was followed by CD. The MRE at 222 nm is reported. (d) Thermal denaturation and renaturation of designed armadillo repeat proteins. From the top, YI_4A, YT_4A, and YC_4A are shown. For comparison, the bottom graph shows the irreversible denaturation of αArm. βArm shows a similar irreversible denaturation (data not shown). The denaturation was followed by CD. The values of MRE at 222 nm were normalized by setting the initial and the final values of the denaturation curves as 0 and 1, respectively. (e) Guanidinium-chloride-induced denaturation of armadillo repeat proteins containing eight internal modules. Comparison of YI_8A, YT_8A, and YC_8A with αArm. The denaturation was followed by CD. The values of MRE at 222 nm were normalized by setting the initial and the final values of the denaturation curves as 0 and 1, respectively. (f) Emission spectra of ANS in the presence of designed armadillo repeat proteins. YI_4A, YT_4A, and YC_4A are compared to αArm and βArm. I- and T-type proteins show fluorescence levels in the same range as natural proteins; in contrast, the fluorescence emission for C-type proteins is significantly higher and increases with the number of repeats. The values without buffer subtractions are shown. αArm was measured in a separate experiment and scaled according to the values of YC_4A present in both sets of experiments. Similar results were obtained with proteins containing two or eight internal repeats.

internal repeats, respectively, have lower midpoints of transition, even when compared with designed proteins with only 4 internal repeats (Table 1). It should be noted that the designed proteins retain a significant percentage of secondary structure at 95 °C and that the thermal unfolding is almost

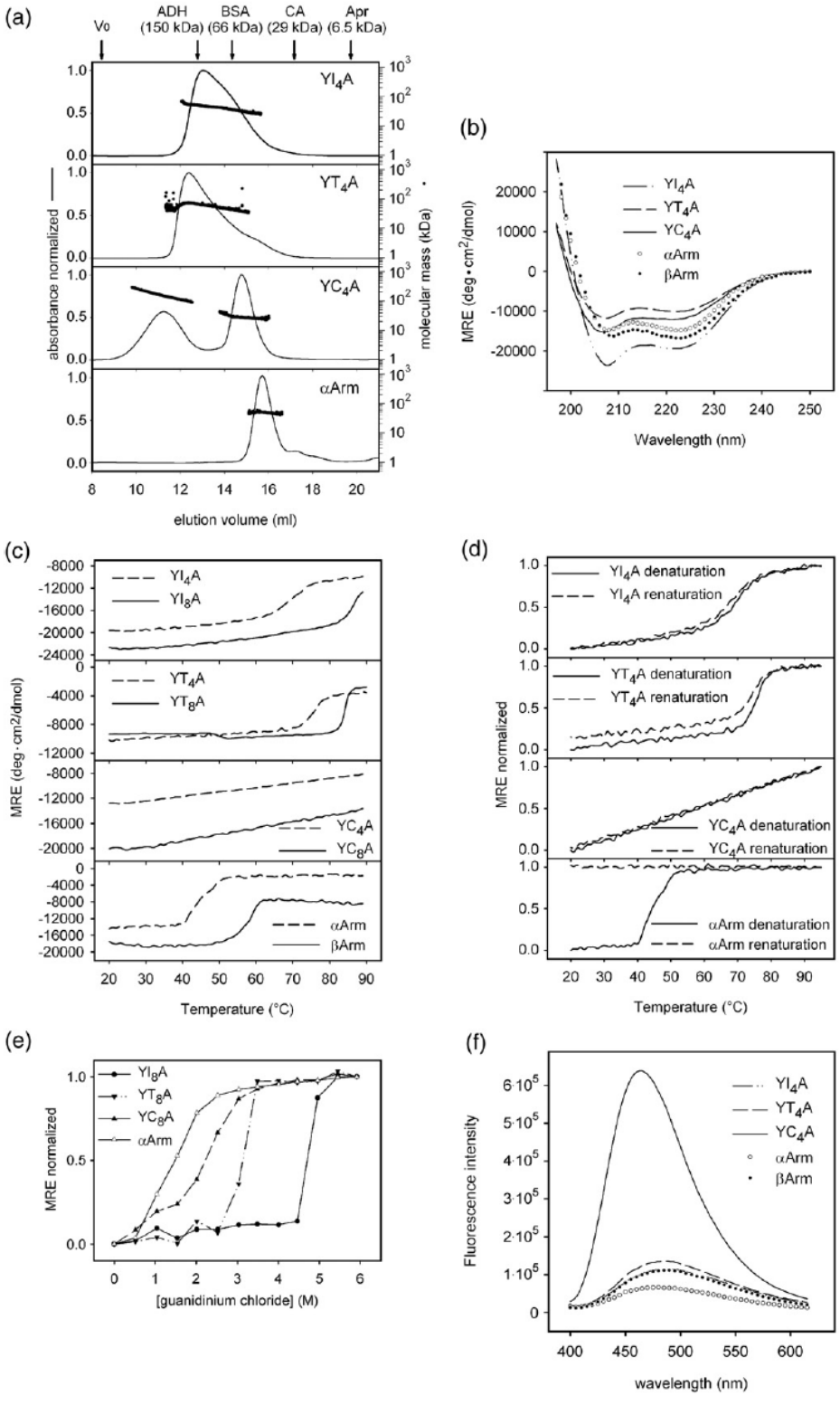


Fig. 5 (*legend on previous page*)

completely reversible, in contrast to natural armadillo proteins that cannot refold after thermal unfolding (Fig. 5d); YT_8A is the only designed armadillo repeat protein whose thermal unfolding is irreversible (data not shown).

We also investigated unfolding induced by guanidinium chloride. A direct comparison between natural and designed armadillo repeat proteins composed of 10 repeats (Fig. 5e) reveals for importin-α1 (αArm), with a midpoint of transition of 1.4 M guanidinium chloride, a lower stability than that for YI_8A and YT_8A, with approximately 4.8 and 3.2 M as midpoints of transition, respectively. YC_8A shows a gradual loss of secondary structure, especially at low concentrations of denaturant, apparently similar to αArm. Data from urea-induced unfolding experiments confirm the gradual loss of secondary structure for C-type proteins with increasing denaturant concentration. Natural armadillo domains show a stable pretransition baseline in unfolding induced by the weaker denaturant urea (data not shown).

The three types of consensus proteins (C-, I-, and T- type) also show a different behavior in 1-anilino-naphthalene-8-sulfonate (ANS) binding experiments. ANS is a fluorescent dye sensitive to the hydrophobic environment.[42] C-type proteins bind ANS strongly, suggesting the presence of an accessible hydrophobic core, while I- and T-type proteins show ANS binding in the same low range as the natural armadillo repeat proteins (Fig. 5f).

Thermal and guanidinium-induced denaturation and ANS results indicate that I- and T-type proteins share many characteristics with proteins with stable folds. Based on MALS data, however, I-type proteins are mainly present as dimers. T-type proteins show even higher deviations of elution behavior in SEC, and remarkably, the helical content does not seem to be significantly affected by the number of internal repeats, in contrast to the T_m value. C-type proteins, though monomeric, are characterized by strong ANS binding, an elution volume smaller than expected for a monomeric protein in SEC, and lack of cooperativity in thermal and chemical denaturation. These features, similar to some extent to the properties of molten globules,[43] indicate that the C-type proteins are probably not folded in a well-packed conformation, even though the expected secondary structure is detected by CD. Nonetheless, we chose the C-type proteins as the basis for our further investigations.

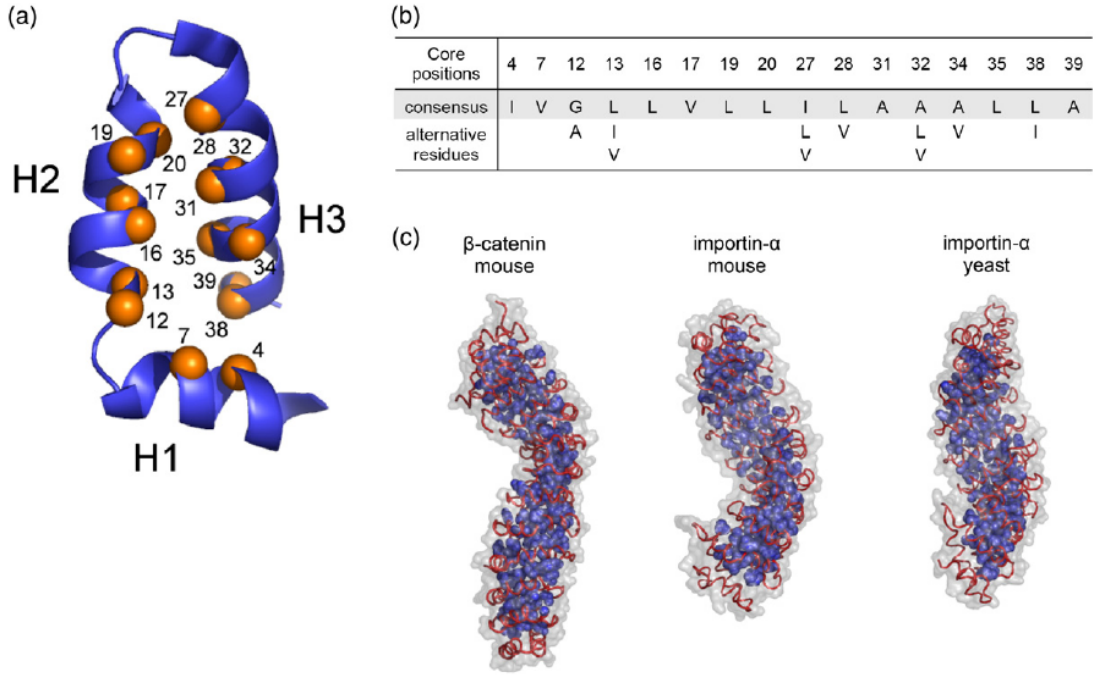
Consensus design improvement: Substitutions in the hydrophobic core

Due to the lack of conserved interrepeat hydrogen bonds and salt bridges, the tertiary structure of natural armadillo repeat proteins holds together mainly through nonpolar interactions. If the packing is not ideal, alternative conformations may become accessible. As a consequence, the molten-globule-like features of C-type proteins could be due to nonoptimal packing of the hydrophobic core.

The modular architecture of designed armadillo repeat proteins suggests that the computational search for a sequence leading to stable packing of the hydrophobic core might be achievable by considering a single repeat. However, the repeat can assume its correct conformation only in the context of a complete protein. It was, therefore, necessary to use the known structures of natural armadillo domains (comprising 400 to 500 residues) as templates for the sequence search.

The use of available algorithms (self-consistent mean field, dead-end elimination, genetic algorithm, and Monte Carlo search) for structures as large as armadillo domains has so far not been reported, despite recent achievements (reviewed by Butterfoss and Kuhlman[44]); such approaches would be, however, seriously compromised by the computational load and probably not even be possible in the case of dead-end elimination, as suggested by Voigt *et al.*[45] Therefore, we used here a different approach to treat a system of such size: information from sequence alignments was used to reduce the complexity in terms of variable positions and allowed residue types. The selected mutants were ranked according to energy values obtained by rotamer sampling. The method allows, in a simple way, to identify a number of hydrophobic core mutant sequences, which are likely to represent an improvement of the original C-type sequence.

The 16 positions contributing to the hydrophobic core in each repeat (Figs. 3 and 6a) were defined by having a solvent-accessible surface corresponding to less than 5% of the total residue surface, as determined by a probe with 1.4 Å radius. The final choice was made after visual inspection of the structures. The number of mutations was restricted to the most frequently occurring aliphatic amino acids at each position, based on the sequence alignment, while keeping the most conserved positions constant. Using these criteria, only 7 positions out of the 16 forming the hydrophobic core of a single repeat were allowed to vary and to host two or three different residue types (Fig. 6b). Mutants were modeled starting from three different backbones to average the influence of single structures out. Therefore, the structures of three different proteins {mouse β-catenin [Protein Data Bank (PDB) ID 2BCT],[22] yeast importin-α (PDB ID 1EE4),[46] and mouse importin-α (PDB ID 1Q1T[47])} were chosen to generate all the mutants (Fig. 6c). Model structures were constructed by substituting the core positions of every internal repeat with either the residues present in the C-type consensus or the aforementioned mutations (Fig. 6b). The initial rotamer conformations were randomly assigned. The noncore residues of the original structures were kept. In each structure, every repeat of the protein carries the same mutations. Structures corresponding to all the 432 combinations of allowed mutations, including also the set of residues of the original



Core positions	4	7	12	13	16	17	19	20	27	28	31	32	34	35	38	39
consensus	I	V	G	L	L	V	L	L	I	L	A	A	A	L	L	A
alternative residues			A	I V					L V	V		L V	V		I	

Fig. 6. (a) Hydrophobic core positions in a single repeat are indicated with orange spheres and the corresponding numbers. (b) Amino acids, in a single-letter code, allowed at the core positions during the calculations. The original amino acids present in the type C consensus are highlighted in gray. The total number of different combinations in each repeat is 432 ($2^4 \times 3^3$). The number of mutants is also 432 because the same mutation pattern was applied to all repeats in each protein. (c) Armadillo domains used as starting structures for the models of the mutants: murine β-catenin (PDB ID 2BCT) and importin-α from mouse (PDB ID 1Q1T) and *S. cerevisiae* (PDB ID 1EE4). The backbone trace is shown in red, and the protein surface is indicated in gray. The side chains belonging to the hydrophobic core residues, which correspond to the parts allowed to move freely during the simulation, are depicted in blue.

C-type consensus, were generated and subjected to energy minimization.

A sequence of heating–quench cycles (Fig. 7), followed by energy minimization, resulted in a series of structures and corresponding energy values that were used to generate the final ranking of the mutants (Supplementary Table S3). A detailed description of the rotamer sampling procedure is provided in Materials and Methods. Mutants with a hydrophobic core volume lower than the original consensus, calculated with values reported by Chothia,[48] were not included in the final ranking to reduce the number of false positives that might arise due to underpacking of the core (see Discussion).

Gene assembly, expression, and characterization of selected hydrophobic core mutants

Among the 30 top-ranked single repeat mutant sequences, 18 were selected for experimental validation. The best-ranking mutant sequence with low core volume was also selected to challenge the initial choice of a core volume filter during the ranking process (Table 2 and Supplementary Table S3). The influence of mutant repeats on the protein properties was experimentally evaluated in the format of proteins containing four identical internal repeats and Ny and Ca as capping repeats (Fig. 3). The original reference consensus sequence is thus YC_4A. The proteins were named with a progressive number, from mut1 to mut18; mut19 contains the sequence with low core volume.

The assembly of single repeats from oligonucleotides and the stepwise ligations were performed as described above, and the proteins were expressed and purified by IMAC in a single step with yields comparable to those obtained for YC_4A, that is, up to 100 mg/l of bacterial culture.

The experimental comparison was carried out by using CD, SEC, and binding of ANS. All the mutants share a similar CD spectrum with the original consensus but are characterized by a general increase in MRE at 222 nm, indicating a higher percentage of α-helical secondary structure. The increased elution volume of the mutants indicates a higher compactness of the proteins (Fig. 8) and correlates well with a decreased ANS binding. The mutant mut1, being a dimer, represents the only outlier, while all the other mutants are monomers, as indicated by MALS. Some of the core mutants carry additional mutations (indicated in Table 2), which were unintentionally introduced during the gene synthesis. Most of these mutations are located in the loops or at the surface of the helices and, thus, have probably only a small

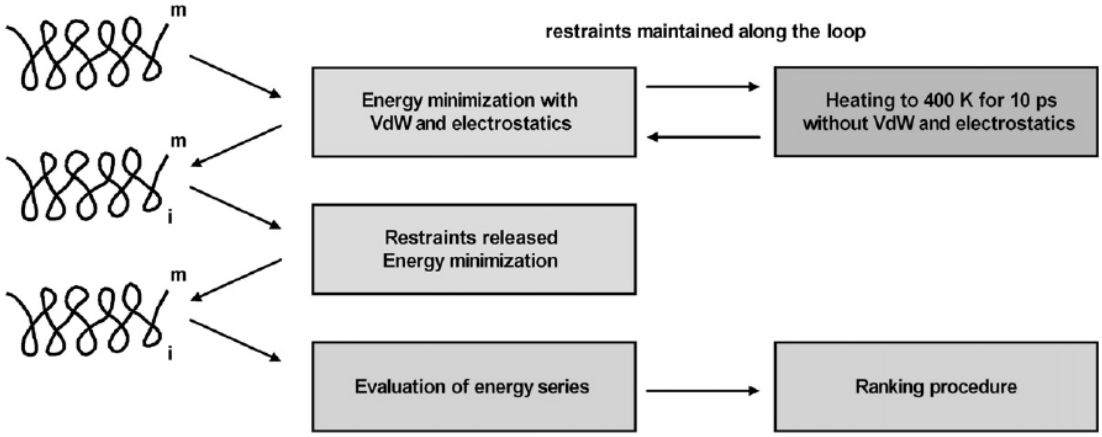


Fig. 7. Schematic diagram of the computational procedure for the evaluation of the hydrophobic core mutants. *m* indicates a particular mutant, and *i* is 1 of the 100 conformations of the mutant *m* obtained after each minimization step in the recursive sampling procedure. VdW, van der Waals interactions.

influence, if any, on the stability of the hydrophobic core; furthermore, they are present only in a single repeat out of four, reducing their overall contribution to protein properties.

Mutants mut2, mut3, mut4, mut7, mut11, mut12, and mut13 showed the best combination of low ANS binding and compactness, as judged by SEC, and were thus selected for further characterization by thermal denaturation. The mutant mut7 shows a significantly increased cooperativity during unfolding, compared to YC_4A and the other mutants (Supplementary Fig. S5).

The internal module corresponding to mut7, which was named M-type, contains three point mutations compared to the initial consensus sequence (Fig. 3). The mutant protein mut7, renamed YM_4A, is a stable monomer at several salt and protein concentrations, such as YC_4A; however, dimer formation of YM_4A was observed at pH 7 at high protein concentrations (5 mg/ml). No sign of precipitation or degradation was detected in protein solutions stored for up to 1 month at 4 °C in the IMAC elution buffer. The values for the biophysical properties examined are reported in Table 1.

The direct comparison of YC_4A and YM_4A is shown in Fig. 9. The [^{15}N,^{1}H]-heteronuclear single quantum coherence (HSQC) NMR spectra of YM_4A were recorded at pH 7, 8, 9, 10, and 11. YC_4A spectra

Table 2. Hydrophobic core of the selected mutants

	Hydrophobic core residues															
	4	7	**12**	**13**	16	17	19	20	**27**	**28**	31	**32**	**34**	35	**38**	39
C-type	I	V	**G**	**L**	L	V	L	L	**I**	**L**	A	**A**	**A**	L	**L**	A
mut1	–	–	**–**	**–**	–	–	–	–	**L**	**–**	–	**L**	**–**	–	**–**	–
mut2	–	–	**A**	**–**	–	–	–	–	**V**	**–**	–	**L**	**–**	–	**–**	–
mut3	–	–	**A**	**–**	–	–	–	–	**L**	**–**	–	**L**	**–**	–	**–**	–
mut4	–	–	**A**	**–**	–	–	–	–	**L**	**–**	–	**–**	**–**	–	**I**	–
mut5	–	–	**A**	**–**	–	–	–	–	**L**	**–**	–	**–**	**V**	–	**–**	–
mut6	–	–	**A**	**–**	–	–	–	–	**L**	**–**	–	**–**	**–**	–	**–**	–
mut7	–	–	**A**	**–**	–	–	–	–	**–**	**–**	–	**L**	**–**	–	**I**	–
mut8	–	–	**A**	**V**	–	–	–	–	**L**	**–**	–	**L**	**–**	–	**–**	–
mut9	–	–	**A**	**I**	–	–	–	–	**L**	**–**	–	**L**	**–**	–	**–**	–
mut10	–	–	**A**	**–**	–	–	–	–	**L**	**–**	–	**V**	**–**	–	**–**	–
mut11	–	–	**A**	**–**	–	–	–	–	**V**	**–**	–	**L**	**V**	–	**–**	–
mut12	–	–	**A**	**–**	–	–	–	–	**L**	**–**	–	**L**	**V**	–	**–**	–
mut13	–	–	**A**	**I**	–	–	–	–	**L**	**–**	–	**L**	**V**	–	**–**	–
mut14	–	–	**A**	**–**	–	–	–	–	**L**	**–**	–	**V**	**V**	–	**–**	–
mut15	–	–	**A**	**–**	–	–	–	–	**–**	**–**	–	**–**	**–**	–	**–**	–
mut16	–	–	**–**	**–**	–	–	–	–	**L**	**–**	–	**L**	**–**	–	**I**	–
mut17	–	–	**A**	**V**	–	–	–	–	**L**	**–**	–	**L**	**V**	–	**I**	–
mut18	–	–	**A**	**I**	–	–	–	–	**–**	**–**	–	**–**	**–**	–	**–**	–
mut19	–	–	**–**	**–**	–	–	–	–	**L**	**–**	–	**–**	**–**	–	**–**	–
I-type	–	–	**A**	**–**	–	–	–	–	**–**	**Q**	–	**L**	**–**	–	**I**	T

The numbers indicate the positions in the single repeat (cf., Fig. 3). The hydrophobic core positions subjected to mutation (12, 13, 27, 28, 32, 34, and 38) are indicated in boldface. The amino acids present at each position are reported as single-letter code. "–" indicates no difference with respect to C-type consensus. As a comparison, in the last row, the sequence corresponding to the I-type consensus is shown. An Ala→Thr mutation occurs in mut6 at position 15 in repeat 3, in mut9 at position 31 in repeat 4, in mut10 at position 12 in repeat 1, and in mut17 at position 15 in repeat 1. mut8 has a mutation Gly→Val at position 42 in repeat 4.

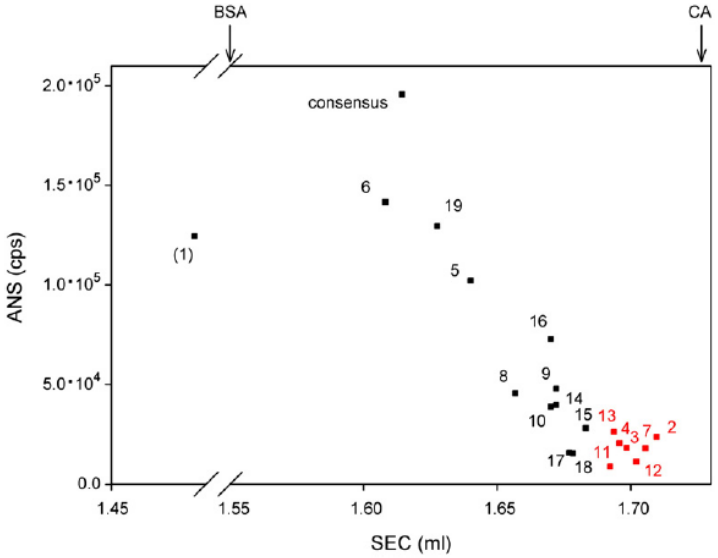


Fig. 8. Experimental evaluation of hydrophobic core mutants: elution volumes in SEC and fluorescence emission upon ANS binding. The numbers refer to the mutants reported in Table 1. *Consensus* indicates the protein containing four C-type internal repeats (YC_4A). All the proteins have a molecular mass of approximately 27 kDa. *mut1* (in parentheses) elutes before the consensus and the other mutants because of its dimeric state. All other mutants were shown to be monomeric by MALS. Peak values from absorbance at 280 nm in SEC and from fluorescence intensity are plotted. Errors in the measurements have been estimated with a subset of six proteins and two different preparations, leading to an average standard deviation of 0.01 ml for SEC and an average percentage error of 4% for ANS fluorescence intensity. As reference, carbonic anhydrase (CA; MW = 29 kDa) and bovine serum albumin (BSA; MW = 66 kDa) elute at 1.73 and 1.55 ml, respectively. The mutants depicted in red were selected for further characterization.

were collected at pH 6, 7, and 8. An increase in pH increases the line broadening of YC_4A but decreases it for YM_4A. Nevertheless, the overall dispersion is conserved for each protein at different pH values (data not shown). The YM_4A spectrum recorded at pH 11 and the YC_4A spectrum recorded at pH 6 are shown in Fig. 10. Amide proton frequencies of YC_4A are generally limited to the random-coil range (7.5–8.5 ppm), whereas many cross peaks of YM_4A are located outside this range. Moreover, the line widths from signals of YC_4A are slightly larger than those from signals of YM_4A. Increased line widths due to conformational exchange processes as well as limited signal dispersion are characteristic features of molten globule states of proteins.[49,50] Although no attempts have been made to assign the $^{15}N,^{1}H$ correlation map, $^{15}N\{^{1}H\}$-nuclear Overhauser enhancement (NOE) data were recorded to characterize internal backbone dynamics[51] and to probe for increased rigidity of YM_4A (data not shown). All detected amide moieties of YM_4A are characterized by $^{15}N\{^{1}H\}$-NOEs larger than 0.6, indicating well-folded segments, whereas for YC_4A, all the values are smaller than 0.3, many of which have negative NOEs, indicating a large flexibility. Thus, the NMR measurements confirm the molten-globule-like characteristics of YC_4A and the folded state properties of YM_4A.

Binding assay as functionality test

YM_4A and YC_4A share with natural importins a considerable number of residues critical for binding to nuclear localization sequences (NLSs), which are the natural ligands of importin-α proteins. Therefore, the designed proteins might retain some binding properties toward NLS. The NLS from the SV40 large T antigen[52] is considered a prototype sequence: it has been extensively studied in the literature and constitutes the reference point for the evaluation of NLS binding.[47]

The NLS from SV40 large T antigen (SPKKKRKVE) was expressed as a fusion protein with phage lambda protein D (pD), biotinylated, and immobilized on NeutrAvidin-coated plates. Being of similar size, the hemagglutinin tag (YPYDVPDYA, here referred to as HA), also fused to protein D, was used as a negative control. ELISA experiments (Fig. 11) reveal that both YM_4A and YC_4A bind specifically to the NLS and that the binding can be competed by a free NLS peptide in solution. However, the unspecific binding of YM_4A to HA and NeutrAvidin is reduced in comparison to YC_4A.

In summary, even though the high concentrations of protein and competing peptide indicate a rather weak affinity, YM_4A was able to specifically recognize the same target as the natural armadillo repeat proteins and to reduce the unspecific binding observed for YC_4A, further validating the design process.

Discussion

Consensus design

Consensus design has been successfully applied in this work to generate designed armadillo repeat proteins. Similar to leucine-rich repeat proteins,[30] but in contrast to ankyrin repeat proteins[27] and tetratricopeptide repeats,[29] different subfamilies can be clearly defined in the case of armadillo repeat proteins, based on sequences and available structures. Out of 42 signature positions, 12 are char-

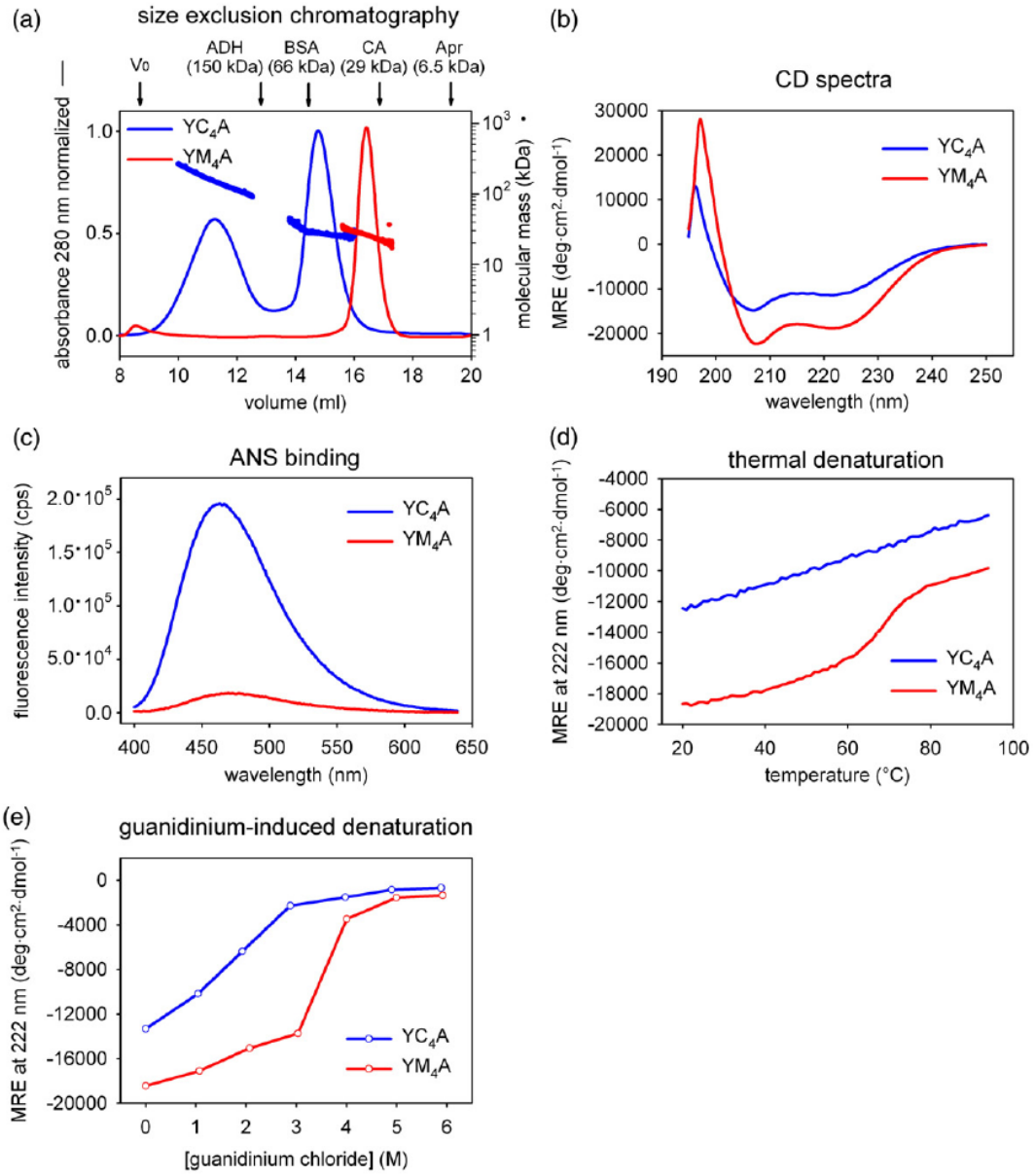


Fig. 9. Comparison between YC_4A, in blue, and YM_4A, in red. SEC (a) was performed with samples directly after IMAC purification. MALS data are also shown. The chromatogram of YC_4A displays one peak corresponding to the monomer (on the right) and one corresponding to oligomeric fractions (on the left). CD spectroscopy (b) shows an increase in ellipticity for YM_4A. ANS binding (c) is drastically reduced for YM_4A to levels typical of natural armadillo repeat proteins; the data shown refer to values after buffer subtraction. Thermal denaturation (d) and guanidinium-induced denaturation (e) indicate the presence of a cooperative unfolding transition, characteristic for native-like proteins, for YM_4A. V_0 indicates the column void volume. Alcohol dehydrogenase (ADH; MW = 150 kDa), bovine serum albumin (BSA; MW = 66 kDa), carbonic anhydrase (CA; MW = 29 kDa), and aprotinin (Apr; MW = 6.5 kDa) were used as molecular weight markers, and the corresponding elution volumes are indicated by arrows.

acteristic for armadillo repeats, but the conservation at other positions is relatively low.[32] To obtain a more reliable and informative consensus, we deemed it necessary to analyze the subfamilies independently. The use of closely related sequences should also improve the self-compatibility between designed repeats. At the time of the initial sequence design, only members of importin-α and β-catenin/plakoglobin subfamilies were known to be peptide binders and had crystal structures available. As a consequence, only the repeats from proteins belonging to these subfamilies were thus chosen for the calculation of the consensus, to avoid interference from other subfamilies of unknown structure that could negatively affect the final sequences. Indeed, the later publication of the structure of plakophilin[53] (a member of the p120 subfamily) revealed an unexpected shape with a pronounced bend in the

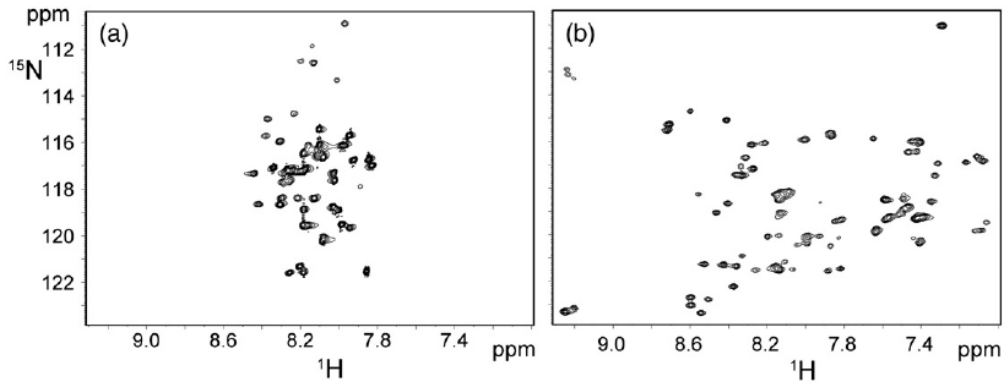


Fig. 10. [^{15}N,^{1}H]-HSQC spectra of designed armadillo repeat proteins: YC_4A (a) at pH 6 and YM_4A (b) at pH 11. Both spectra were recorded at a temperature of 310 K in 20 mM Tris–HCl and 30 mM NaCl. The protein concentration was 0.6 mM.

middle of the domain, supporting, *a posteriori*, the initial choice of sequence restriction to the subfamilies mentioned above.

An overall consensus was, however, also realized to take into account the possible combination of sequences belonging to importin-α and β-catenin/plakoglobin subfamilies. An obvious concern regarding the combination of these two subfamilies was that the overall consensus (type C) might be too similar to the importin consensus (type I) due to the slight overrepresentation of importin sequences in the original set. After the exclusion of the positions involved in binding, highly conserved for functional reasons especially in the importin subfamily and thus preserved also in the overall consensus, the C- and I-type repeats share 74% identity and 82% similarity, while C- and T-types have corresponding values of 70% identity and 87% similarity. The values indicate that the overall consensus is thus not significantly biased toward the importin consensus in the "framework" positions, that is, the positions not responsible for binding. The positions involved in peptide binding will be randomized in the library design and thus do not play a role in these considerations. Nevertheless, despite the similarity between I-, T-, and C-type modules, we always used only one type of consensus modules in every repeat protein tested, to provide a constant interface between the repeats and to be able to correlate the protein properties with the types of modules.

The capping repeats represented a second key point in the protein design. As observed for designed ankyrin repeat proteins,[31,54] capping repeats can dramatically increase *in vivo* folding yield and prevent aggregation. We found that an N-terminal capping repeat derived from yeast importin-α (Ny) and an artificial C-terminal capping repeat (Ca), designed by replacing exposed hydrophobic residues, give the highest expression yield of soluble protein in *E. coli*. Remarkably, we could find a single combination of capping repeats that allowed us to analyze the properties connected to the types of internal modules.

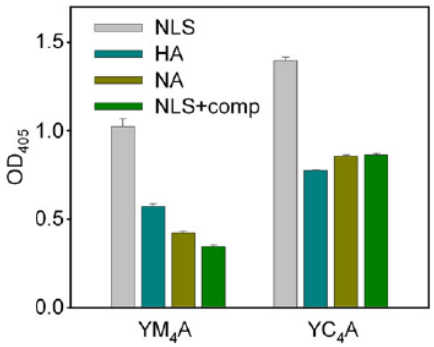


Fig. 11. ELISA of YM_4A and YC_4A. YM_4A binds specifically to immobilized SV40 large T antigen NLS. Immobilized hemagglutinin tag peptide (HA) and NeutrAvidin (NA) are negative controls. Binding to NLS can be competed by addition of NLS peptide (SPKKKRKVE) in solution at a concentration of 10 μM (NLS+comp). Experiments were performed in duplicate with YM_4A and YC_4A at a concentration of 1 μM.

Protein properties

Data from the artificial repeat proteins previously designed[28–30] indicate that biophysical properties often correlate with the number of internal repeats. Indeed, this behavior was also observed for designed armadillo repeat proteins.

I-type proteins

I-type proteins show that helical secondary structure content, thermal stability, and resistance to guanidinium-induced denaturation increase with the number of repeats, pointing in the same direction as data from other artificial repeat proteins.[27,29,30,55] A helical content of approximately 80% for YI_4A and YI_8A (Table 1) corresponds to the expected theoretical value from the design and is even higher than the values observed for natural armadillo domains. Low ANS binding and clearly defined transitions in thermal and guanidinium-

induced denaturation indicate that the I-type module can lead to native-like molecules, and the elevated midpoint of denaturation points toward a superior thermodynamic stability compared to natural proteins. At the same time, thermal denaturation is almost completely reversible. The thermal denaturation was employed here as a qualitative method to assess the stability of the designed proteins and to compare them to their natural counterparts. A detailed thermodynamic analysis requires, however, further investigation of the folding mechanism, which is probably more complex than a simple two-state transition and possibly also described by the Ising model as in the case of other designed repeat proteins.[56,57]

I-type proteins could, thus, be good candidates as scaffold for peptide-binding molecules. However, their predominant dimeric state constitutes a disadvantage during selection and characterization of binding properties due to possible avidity effects. Even considering that the I-type proteins are dimers, the SEC data indicate an elution volume still larger than expected, which could be interpreted as a result of an elongated shape. It is noteworthy that natural armadillo proteins do not show a higher-than-expected apparent mass in gel filtration (Table 1).

T-type proteins

T-type proteins share several native-like characteristics with I-type proteins, such as the presence of a compact hydrophobic core inaccessible to solvent, as suggested by ANS binding levels that are as low as those of natural armadillo repeat proteins, and the transitions observed in thermal and guanidinium-induced denaturation. The reversibility of thermal denaturation in T-type proteins is, however, less complete than that in I-type proteins and completely lost in the case of YT_8A. The helical content of T-type proteins is approximately independent of the number of repeats and generally lower than that in natural armadillo repeat proteins. The gel filtration results are, however, similar to I-type proteins, with apparent molecular masses higher than expected by more than a factor of 3 on average, already taking the dimeric state into account (Table 1). Due to the native-like properties of T-type proteins, the increase in hydrodynamic radius could be interpreted as an effect of a rodlike shape. Despite the different behavior in gel filtration, T-type proteins are therefore more similar in their biophysical characteristics to native armadillo proteins than to an idealized scaffold. When applying a strategy of protein assembly from preselected modules, which represents one of the aims of a general modular peptide binder, the scaffold properties should ideally change in a regular and predictable way, when adding modules, without altering the general characteristics. However, this is not observed for YT_8A, where the reversibility in thermal unfolding is completely lost.

C-type proteins

C-type proteins, in contrast to the other designed armadillo proteins, do not show a clear transition in thermal or guanidinium-induced denaturation but a gradual loss of secondary structure, and ANS binding results indicate the presence of an accessible hydrophobic core. Thus, C-type proteins are probably not completely folded but rather are in a molten-globule-like state. The secondary structure is present, as indicated by CD, but the proteins retain a high level of flexibility due to the lack of a fixed tertiary structure. The high apparent molecular masses observed in gel filtration, where MALS indicate monomeric states, might then be interpreted as a consequence of the intrinsic flexibility of the polypeptide chain. The molten-globule-like characteristics of C-type proteins represent a serious limitation in library generation, where framework stability and tolerance to mutations are desired. From the point of view of the design, however, the observation of a molten-globule-like state for armadillo repeat proteins built from overall consensus (type C) modules could suggest either an insufficient stability of each repeat or an inadequate interaction between them, supporting the initial choice of restricting the design to specific subfamilies.

Molten globule stabilization

The initial consensus-based approach led to stable dimers (I- and T-types) or molten-globule-like monomeric proteins (C-type). The further possible design steps to obtain a stable monomer were either the disruption of the interaction in the dimer or the stabilization of the C-type proteins. However, no information was available concerning the dimerization interface and the residues involved; both surface interaction and domain swap were conceivable as mechanisms of dimer formation. The improvement strategy would thus have to involve systematic point mutations of several single residues and combination of residues, with the risk that the disruption of the dimer will simply lead to a stability loss or even a molten globule state. For disrupting a dimer, the improvement strategy would have to consist of systematic mutations of single residues or combination of residues without a structural hint to select mutations.

We chose instead an alternative approach, focused on the stabilization of the hydrophobic core of the molten-globule-like C-type proteins, using a computational approach. As mentioned above, the molten-globule-like state suggested inadequate inter- or intra-repeat interactions at the hydrophobic core level and, hence, insufficient packing of the core. Our results show that the introduction of only three point mutations in the hydrophobic core of C-type repeats was sufficient to convert a molten-globule-like protein with four internal repeats to a stable conformation. This strongly argues that the packing of the hydrophobic core was indeed the critical parameter for obtaining a stable fold.

The underpacking of the core may be one of the reasons for the molten globule behavior[58] of the C-type proteins. Two of the mutations (Gly to Ala at position 12 of the repeat and Ala to Leu at position 32) increase the calculated volume of the hydrophobic core, bringing it close to the average value of natural repeats. These residues are also the most common among the 50 highest-ranking sequences with a frequency of 72% for Ala12 and 50% for Leu32. The third mutation (Leu to Ile at position 38) probably reduces the local flexibility by limiting the number of available rotamers. Such a restriction can help to lock the hydrophobic core in a unique conformation, and this positive effect could overcome the disadvantage of having a residue with low helical propensity, such as isoleucine. However, as observed for several mutants, the contributions from the single residues are not additive and the core packing is the result of a particular combination of residues.

Strikingly, the M-type repeat has, among the 432 mutants screened *in silico*, the core sequence closest to the I-type repeat (Table 2). The only two core residues that differ between M- and I-type repeats (Gln28 and Thr39) were not included in the set of possible mutations in our computational approach. The protein YI_4A, derived from I-type modules, shows characteristics very similar to YM_4A, apart from its dimeric state. Therefore, the particular core sequence obtained for both types of repeats represents a reliable solution for core packing, considering that it has been obtained by consensus design (for I-type) and simulation (for M-type). The hydrophobic core is probably rather stable, and we may speculate that the dimerization observed for I-type proteins takes place most likely via surface interaction instead of domain swap. The introduction of surface point mutations could then possibly lead to the formation of stable monomers.

YM_4A

The observed biophysical characteristics indicate that YM_4A represents a significant improvement of the original consensus sequence. YM_4A is almost as compact as globular proteins with similar molecular weight, as judged from elution volumes in SEC, and only marginally binds ANS, with values in the range observed for natural armadillo repeat proteins. The thermal and guanidinium-induced denaturation curves have sigmoidal profiles, indicating the presence of a cooperative unfolding, a hallmark of natural globular proteins.

NMR

NMR spectra provide further indications of the folded structure of YM_4A. Due to the repetitive nature of the sequence, it is *a priori* not clear how many peaks should be expected, but, even in the absence of specific assignments and considering the effects of symmetry, most of the peaks are probably present. However, Gly residues, usually observed in a characteristic region of the correlation map, are missing, most likely due to the highly accelerated amide proton exchange at pH 11 for residues outside the regions of secondary structure. Nevertheless, the presence of most peaks at the elevated pH indicates that the majority of amide moieties are protected from solvent access.

Although it was not possible to assign the spectra, the $^{15}N\{^1H\}$-NOE data indicate that almost all peaks in the proton–nitrogen correlation spectrum correspond to amide moieties with motional properties similar to those of residues from stably folded secondary structural elements. Hence, the NMR measurements suggest that YM_4A at pH 11 can be considered as a well-folded globular protein, whereas YC_4A shows characteristics of a molten globule. YM_4A, at pH lower than 10, displays broader lines, without affecting the signal dispersion, indicating that under those conditions, good side-chain packing is probably disturbed by the presence of an ionized group. A large range of pH values was also tested for YC_4A, yet without leading to any improvement in the dispersion of the signals in $[^{15}N,^1H]$-HSQC spectra or narrowing of the line width. Hence, electrostatic interactions are not dominating the molten globule properties of YC_4A. These observations rather indicate that subtle effects of side-chain packing are involved and that proper side-chain packing is achieved only in YM_4A, which presumably requires a neutral state of one group that is charged at neutral pH but uncharged at pH 11. As the lines are sharper at basic pH, lysine residues are the candidates for causing this effect, because of the possible repulsive interaction with the lysines in the neighboring repeats when both are charged, as observed in the molecular models.

Peptide binding

The binding to the SV40 large T antigen NLS observed in ELISA confirms the interpretation of the biophysical data. Unspecific binding has been reduced in YM_4A, compared to the original molten-globule-like YC_4A, as observed in binding to NeutrAvidin and to the hemagglutinin tag and in the competition experiment.

Even though no design effort was made in the present work for binding to a target peptide, YM_4A and YC_4A do show a weak but specific binding, indicating a correct disposition of the residues involved in interactions with the peptide. Glu30, Trp33, and Asn37 in the M-type repeat correspond to the residues responsible for binding to NLS in natural importin-α proteins. These residues are present in YC_4A and YM_4A due to the high conservation in the importin-α sequences, which were used in the original consensus design. The competition with soluble peptide strongly suggests the presence of specific interactions rather than a merely electrostatic binding phenomenon.

Further experiments will be needed to clarify the binding of consensus-designed armadillo repeat proteins. Nevertheless, the results already achieved indicate a correct structure. Armadillo repeat proteins based on M-type repeats can thus be used as scaffold for library generation and selection.

Evaluation of the computational method

A rotamer sampling method was chosen to identify, from a large pool of candidates, armadillo repeat proteins with improved core packing. The approach was devised for use with large proteins, up to 500 residues in our case. Despite recent advances,[59] such complexity is still not easily treatable by the available methods for core repacking, which proceed through a cycle of mutation, selection of residue conformation, and energy minimization. In terms of computational load, the search for a sequence with minimal energy is highly demanding, or even not affordable at all, for large proteins. In contrast, a simple evaluation of the potential energy of protein models after energy minimization is rather unreliable. The introduction of point mutations in the hydrophobic core requires the rearrangement of the core side chains to optimize the core packing. This task is, however, not fulfilled by a simple energy minimization, especially when the energy barrier between rotamers is too high to be overcome and only the nearest local minima for the side chains are reached (e.g., for tightly packed side chains). However, our aim was not to find the conformation at the global minimum but to estimate the packing efficiency of given mutants. A random sampling, helped by the partial removal of the energy barriers and followed by statistical analysis, is thus a feasible procedure for evaluating the packing of each mutant protein.

Though being a simplified approach, it was still necessary to reduce the complexity of the system. The choice of candidate sequences was based on information derived from sequence alignment and interactions in the structures. This approach is not exhaustive, but it restricts the search space to the most promising mutants according to criteria that are independent of the computational method.

Nevertheless, some further restrictions were required to keep the computational load within reasonable boundaries. Calculations without solvation terms are computationally less expensive and can be applied in our case, where the mutations correspond only to aliphatic–aliphatic substitutions in the hydrophobic core and are not expected to influence solute–solvent energy contributions. When restraining backbone atoms, the influence of the core mutations on the surface electrostatics is negligible and the electrostatic contribution to the potential energy does not vary upon mutation.

A second crucial issue was the choice of three different armadillo structures as starting points for the generation of the mutant models to avoid a result biased by the use of a single structure. Additionally, the introduction of multiple backbone templates and backbone flexibility has already been shown to improve the quality of the sequence search.[60,61] In the present work, we did not attempt to build a new backbone including the information from multiple structures, but we allowed "flexibility" by using three starting structures and fixing the coordinates of the backbone atoms using harmonic restraints. The use of three structures per mutant also helped to enhance the signal-to-noise ratio of the ranking, as a high rank for all three armadillo structural backgrounds is generally required to obtain a high overall rank.

Sequences with low core volume were excluded from the final ranking to decrease the number of false positives in the pool selected for experimental characterization (Supplementary Table S3). One example of such a low core volume sequence is the original C-type consensus, which is highly ranked; its core volume was set as the lower volume threshold for the discrimination of the mutants. On a fixed backbone, a reduced core volume allows side chains to assume nearly ideal values of bond lengths, angles, and dihedrals, leading to a significant reduction of the total energy and to an artificially high rank. An increase in flexibility of the backbone could also be the source of artifacts: the cavities in the hydrophobic core of low volume mutants can be compensated by compressing the backbone structure and bringing the side chains close enough to take advantage of the van der Waals interactions. However, a reduction of the backbone flexibility could be detrimental for mutants slightly more overpacked than the natural structures, which would not be able to reach low energy values without backbone adjustments.

Homology models of C-type proteins, based on the armadillo crystal structures, indicated the likely presence of small cavities in the hydrophobic core, suggesting that underpacking is one of the possible reasons for the molten globule state. It is thus unlikely that proteins with core volume lower than the original C-type consensus can provide better packing. Furthermore, mut19, the highest-ranked mutant with low core volume, displays ANS binding and SEC properties closest to the original overall consensus sequence, confirming the validity of our selection filter based on core volume. No threshold level was set for possible overpacking cases: the maximal value of core volume among the considered mutants was still in the range of the average repeat volume calculated for the reference structure of murine importin-α (PDB ID 1Q1T; Supplementary Table S3).

On the experimental side, the use of ANS binding and SEC to discriminate mutants is rather qualitative but can represent an efficient and relatively fast method for screening. A good overall indication of the quality of our method is given by the fact that all the monomeric mutants analyzed show an improvement compared to the original overall consensus.

Conclusions

This work focused on the generation of designed armadillo repeat proteins for the construction of a general modular peptide-binding scaffold. An initial consensus-based design led to well-expressed and stable but dimeric proteins or molten globules. A stable, well-expressed monomeric protein was obtained using a force field-based approach for the stabilization of the hydrophobic core of the molten globule variant.

In a library perspective, a monomeric protein allows a better evaluation of the binding properties, without the influence of possible avidity effects, which can be critical in the discrimination between similar target peptides. The mutations to be introduced to generate a library will only affect surface residues, leaving the hydrophobic core untouched except for one position (position 4 may contribute to both the hydrophobic core and the binding site). Therefore, the favorable characteristics of the designed proteins will probably be kept for most library members and selected specific binders.

Materials and Methods

Sequence analysis and modeling

SMART†,[33,34] Swiss-Prot‡,[26] and PDB§[62] were used as the starting databases for our analysis. GCG (Wisconsin Package Version 10.3, Accelrys Inc., San Diego, CA), BLAST‖,[63,64] and ClustalW¶[35] were used for sequence retrieval and alignment. Structure analysis and modeling were performed with Swiss-Pdb Viewer[a],[65] MOLMOL[b],[66] PyMOL[c] (DeLano Scientific LLC, San Francisco, SA), and INSIGHT II (Accelrys Inc.). Vector NTI (Invitrogen) was used for vector and oligonucleotide design.

General molecular biology methods

Unless stated otherwise, experiments were performed according to Sambrook and Russell.[67] Vent Polymerase (New England Biolabs, USA) was used for all DNA amplifications. Enzymes and buffers were from New England Biolabs or Fermentas (Lithuania). The cloning and production strain was *E. coli* XL1-blue (Stratagene, USA). Competent cells were prepared according to Inoue *et al.*[68] The *E. coli* strain M15 (Qiagen, Germany), containing the plasmid pREP4, was used for the production of ^{15}N-labeled proteins for NMR experiments. The cloning and protein expression vectors were pQE30 (Qiagen, Switzerland) and pPANK (GenBank accession number AY327140). From this, the vector pPANK-NyCa was constructed by cloning of the capping repeats Ny and Ca. pPANK-NyCa contains the BsaI and BpiI restriction sites between the capping repeats for cloning purposes. Note that the inserts were constructed with a double stop codon (see Supplementary Fig. S3). pPANK-NyCa was used to clone the internal repeats for N8C and core mutant proteins. pQE30 and derivatives such as pPANK carry an $MRGSH_6$ tag at the N-terminus of the proteins. The DNA sequences corresponding to the NLS and HA peptides were inserted in the vector pAT223 (GenBank accession number AY327138) and expressed as fusion proteins to pD. The produced proteins consist of N-terminal Avi tag, pD, His_6 tag, and the peptide at the C-terminus. The plasmid pBirAcm (Avidity, USA), encoding *E. coli* biotin-protein ligase BirA, was used for *in vivo* biotinylation of pD peptides.

Cloning of designed armadillo repeat proteins

Oligonucleotides were purchased from Microsynth AG (Balgach, Switzerland). A complete list of all oligonucleotides used is given in Supplementary Table S1. An approach similar to the one described by Binz *et al.*[27] was adopted for gene assembly (Supplementary Fig. S2). All single repeat modules were assembled from oligonucleotides by assembly PCR. The single modules of the core mutants were assembled using the combinations of oligonucleotides indicated in Supplementary Table S2. As an example, for the C-type consensus, pairs of partially overlapping oligonucleotides (1–2, 3–4, and 5–6) were annealed and the double strand was completed by PCR. Then, 2 μl from these PCR reaction mixtures was combined as template for a second assembly reaction in the presence of oligonucleotides 1 and 6. All the oligonucleotides were used at a final concentration of 1 μM. The annealing temperature was 47 °C for the first reaction and 50 °C for the second. Thirty PCR cycles were performed with an extension time of 30 s. The same procedure was applied for the other internal and capping repeats. Only four oligonucleotides were used for the N-terminal capping repeats. BamHI and KpnI restriction sites were used for the direct insertion of the modules into the plasmid pQE30. The single modules were PCR amplified from the vectors, using external primers pQE_f_1 and pQE_r_1 (Qiagen, Switzerland). Neighboring modules were digested with the type IIS restriction enzymes BpiI and BsaI and directly ligated together. The genes coding for the whole proteins were assembled by stepwise ligation of the internal and capping modules. BamHI and KpnI restriction sites were used for insertion of the whole genes into the vector pQE30 and the plasmids were sequenced. For pD-peptide fusions, oligonucleotides encoding both strands of the peptide sequences and containing the restriction sites for BamHI and HindIII were mixed and heated to 95 °C for 10 min and then cooled to 4 °C to allow annealing of the two strands. The double-stranded DNA fragments were subsequently digested with BamHI and HindIII and ligated into the plasmid pAT223.

Natural armadillo domain constructs

The armadillo domain of mouse β-catenin (βArm; residues 150–665) was amplified from the cloned β-catenin gene[22] (a generous gift from W.I. Weis, Stanford University, USA) using oligonucleotides AcatFOR and AcatREV, digested with BamHI and KpnI, and inserted into pQE30. The armadillo domain of human importin-α1 (αArm;

‡ http://www.expasy.org
§ http://www.pdb.org
‖ http://www.ncbi.nlm.nih.gov/blast
¶ http://www.ebi.ac.uk/clustalw
[a] http://www.expasy.org/spdbv/
[b] http://hugin.ethz.ch/wuthrich/software/molmol/
[c] http://pymol.sourceforge.net

residues 83–505) was amplified from a vector containing the importin gene (named importin-α5 in the original publication,[69] a generous gift from M. Köhler, Ostseeklinik Damp, Germany) using oligonucleotides IMAF5 and IMAR5, digested with BamHI and KpnI, and inserted into pQE30. Both proteins carry an N-terminal $MRGSH_6$ tag, as do the designed armadillo repeat proteins.

Protein expression and purification

E. coli XL1-blue cells were transformed with the respective plasmid and grown in LB medium containing 1% (w/v) glucose and 50 μg/ml of ampicillin at 37 °C with vigorous shaking. Expression was induced by IPTG (final concentration of 0.5 mM) when the culture reached OD_{600} = 0.6. After 3 h of expression, cells were harvested by centrifugation. For *in vivo* biotinylation of pD peptides that contain an N-terminal Avi tag, cells were cotransformed with pBirAcm and pAT223 (carrying the pD-peptide constructs) and grown in medium containing 30 μg/ml of chloramphenicol and 50 μg/ml of ampicillin. Before induction with IPTG, biotin was added to the medium to a final concentration of 50 μM, according to Cull and Schatz.[70]

Protein purification was performed at 4 °C. Cells were resuspended in 50 mM Tris–HCl and 500 mM NaCl (pH 8.0) and lysed in a French pressure cell (SLM Instruments, USA) at a pressure of 1200 psi. The lysis mixture was further homogenized by sonication (Branson, USA). Insoluble material was pelleted by centrifugation at 20,000*g* for 30 min. The supernatant was purified by IMAC with Ni–NTA material (Qiagen), equilibrated with buffer containing 50 mM Tris–HCl, 500 mM NaCl, 10% (v/v) glycerol, and 20 mM imidazole (pH 8.0). Columns were washed extensively with the equilibration buffer and then proteins were eluted with an elution buffer identical with the equilibration buffer but also containing 250 mM imidazole. β-Catenin was expressed and purified under the same conditions.

For importin-α1, the expression was carried out at 25 °C for 6 h and the cell pellet was resuspended in lysis buffer containing 50 mM Tris–HCl, 500 mM NaCl, 10% glycerol, 5 mM β-mercaptoethanol, and 10 mM imidazole (pH 8.0). IMAC purification was performed as indicated above using the same buffers with the addition of 5 mM β-mercaptoethanol. Samples were then dialyzed overnight against 50 mM Tris–HCl and 2 mM DTT (pH 8.0) and applied to a POROS HQ anion-exchange column, equilibrated with running buffer (50 mM Tris–HCl, pH 8.0), using the BioCAD 700 E Perfusion Chromatography Workstation (Applied Biosystems, Germany). The column was then washed with 50 mM Tris–HCl and 20 mM NaCl (pH 8.0), and the samples eluted with a gradient from 20 mM to 1 M NaCl. Protein size and purity were assessed by 15% SDS-PAGE, stained with Coomassie PhastGel Blue R-350 (GE Healthcare, Switzerland).

The expected mass of all the studied proteins was confirmed by mass spectrometry. Protein concentrations were determined by absorbance at 235 and 280 nm using molecular masses and extinction coefficients calculated with the tools available at the ExPASy proteomics server‡ and by the bicinchoninic acid assay (Pierce).

SEC and MALS

Analytical SEC was carried out either on an Ettan LC system using a Superdex 200 PC 3.2/30 column (flow rate 70 μl/min) or on an ÄKTA explorer chromatography system using a Superdex 200 10/30 GL column (flow rate, 0.5 ml/min) (GE Healthcare). Phosphate buffer (50 mM phosphate and 150 mM NaCl, pH 7.4) and two Tris-based buffers (20 mM Tris–HCl and 50 mM NaCl, pH 8.0, or 50 mM Tris–HCl and 500 mM NaCl, pH 8.0) were used. The armadillo domain of β-catenin was soluble only at 150 or 500 mM salt concentration. The armadillo domain of importin-α1 was analyzed in phosphate buffer (50 mM phosphate, 500 mM NaCl, and 5 mM DTT, pH 7.4). The core mutants were analyzed in buffer containing 20 mM Tris–HCl and 50 mM NaCl, pH 8.0. MALS measurements were performed with a miniDAWN light-scattering detector and an Optilab refractometer (Wyatt Technologies, USA) coupled to the ÄKTA system. Molecular weight estimates were calculated using the ASTRA 4.73.04 software package (Wyatt Technologies).

CD spectroscopy

CD measurements were performed on a Jasco J-810 spectropolarimeter (Jasco, Japan) using a 0.5-mm cylindrical thermocuvette. CD spectra were recorded from 190 to 250 nm with a data pitch of 1 nm, a scan speed of 20 nm/min, a response time of 4 s, and a bandwidth of 1 nm. Each spectrum was recorded three times and averaged. Measurements were performed at 20 °C. The CD signal was corrected by buffer subtraction and converted to MRE. Heat denaturation curves were obtained by measuring the CD signal at 222 nm with temperature increasing from 20 to 95 °C (data pitch, 1 nm; heating rate, 1 °C/min; response time, 4 s; bandwidth, 1 nm). Data were processed as described above. Guanidinium-induced denaturation measurements were performed after overnight incubation at 20 °C with increasing concentrations of guanidinium chloride (99.5% purity, Fluka), and the data were collected and processed as described above. Measurements of designed armadillo repeat proteins were performed in 20 mM Tris–HCl and 50 mM NaCl (pH 8.0). CD spectra and denaturation curves of the armadillo domain of β-catenin were measured in 50 mM Tris–HCl and 500 mM NaCl (pH 8.0). CD spectra and denaturation curves of importin-α1 were measured in 50 mM phosphate, 500 mM NaCl, and 5 mM DTT (pH 7.4). CD spectra were analyzed using CDpro.[41] Among the algorithms available in CDpro, CDSSTR was chosen for the analysis, with the reference protein set SDP48 (IBasis = 7).

ANS binding

ANS fluorescence was measured using a PTI QM-2000-7 fluorimeter (Photon Technology International, USA). The measurements were performed at 20 °C in 20 mM Tris–HCl, 50 mM NaCl, and 100 μM ANS (pH 8.0) using purified proteins at a final concentration of 10 μM. ANS binding to the armadillo domain of β-catenin was measured in 50 mM Tris–HCl, 500 mM NaCl, and 100 μM ANS (pH 8.0) to avoid possible aggregation problems. The emission spectrum from 400 to 650 nm (1 nm/s) was recorded with an excitation wavelength of 350 nm. For each sample, three spectra were recorded and averaged.

Rotamer sampling of hydrophobic core mutants

A computational approach at the atomic level of detail was used to optimize the hydrophobic core. The approach

uses cycles of energy minimization and heating by molecular dynamics to sample favorable arrangements of the buried side chains and estimate the packing efficiency of residues in the hydrophobic core of a given mutant. The number of possible mutations in each repeat is 432 (Fig. 6b). Three x-ray structures were chosen as starting models to improve sampling: importin-α from *S. cerevisiae* (PDB ID 1EE4[46]) and mouse (PDB ID 1Q1T[47]), consisting each of 8 internal repeats and 2 capping repeats, and murine β-catenin (PDB ID 2BCT[22]), which consists of 10 internal repeats and 2 capping repeats. The original capping repeats of the three structures were substituted with Ny and Ca capping repeats (Fig. 3) in the models. Each mutation was modeled by deleting the side chains at the core positions of each repeat and substituting them with the new side chains with random rotamer conformations; the resulting structure was minimized to eliminate clashes. Three models (from the three initial structures) were prepared for each of the 432 combinations of allowed mutations. All the repeats in each model were designed to have the same mutation pattern.

The extended atom approximation (param19) of the CHARMM force field[71] with a distance-dependent dielectric function was used for both energy minimization and heating by short molecular dynamics runs. All the side-chain atoms not directly in contact with core residue atoms (i.e., those more than 5 Å away, in the initial conformation, from any atom of the 16 core residues of each repeat) and all the backbone atoms were restrained using a harmonic potential with a force constant of 1.0 kcal·mol^{-1}·Å^{-2}. As a consequence, only the side-chain rotatable bonds of the core residues were fully flexible. The system was further minimized in the presence of the harmonic potential. A heating–quench protocol was iterated 100 times for each of the 432 mutants and each of the three protein models (Fig. 7). The first step was a 10-ps heating to 400 K, omitting the nonbonding energy terms (i.e., van der Waals and electrostatics). The second step was a minimization including all energy terms. The aim of the heating phase was to shuffle the flexible side-chain rotamers. The absence of nonbonding energy terms and the high temperature granted a more efficient exploration of the energy landscape. After the heating step, the minimization was used to reach the nearest minimum of the total potential energy. The coordinates were stored at the end of each minimization, and a total of 100 conformations were generated for each mutant. These conformations were further minimized without the aforementioned restraints, and the potential energy was evaluated for ranking.

The CHARMM potential energy is:

$$E = E_{\text{bonding}} + E_{\text{vdw}} + E_{\text{elec}} \quad (1)$$

where E_{bonding} is the sum of bond, angle, improper, and dihedral potential terms; E_{vdw} is the van der Waals energy; and E_{elec} is the coulombic energy. The E_{elec} term was neglected for ranking, because it is insensitive to aliphatic-to-aliphatic mutations in the extended atom representation. Moreover, because of the restraints, the restricted flexibility of the backbone polar groups results in a noisy coulombic energy. Therefore, a reduced potential energy was used for ranking. For each starting structure, the energy value for the conformation i of mutant m is:

$$E_i^m = E_{i,\text{bonding}}^m + E_{i,\text{vdw}}^m \quad (2)$$

As each conformation has a different potential energy, median, first percentile (of most favorable values), and minimum energies were extracted from the energy series of the 100 conformers to characterize each mutant: these values were used to make three independent ranks. At the end of this procedure, for each of the three initial structures, three rank numbers (corresponding to median, first percentile, and minimum ranking) were assigned to each mutant, and these nine rank numbers were summed. Finally, this sum was used for the overall rank of the mutant (Supplementary Table S3). The combination of multiple structures and different scoring criteria (i.e., median, first percentile, and minimum) was used to take into account, in an approximate way, the limited sampling. The central processing unit time required for each starting model of a mutant was approximately 5 h for importin structures and 7 h for catenin structures on a single processor of a 2800-MHz Opteron dual core. The total calculation time of approximately 8000 h was distributed over 150 central processing units.

NMR

Proteins for NMR studies were produced using *E. coli* strain M15 (Qiagen) containing the plasmid pREP4 growing in minimal medium with ^{15}N-labeled ammonium chloride as the only nitrogen source. The medium was supplemented with trace metals, 150 μM thiamin, and 30 μg/ml kanamycin. Expression and purification by IMAC and gel filtration were performed as described. The buffers used for NMR measurements contained 20 mM deuterated Tris–HCl and 30 mM NaCl (pH values of 6, 7, 8, 9, 10, or 11). YC_4A and YM_4A were concentrated to 0.6 mM for NMR measurements.

Proton–nitrogen correlation maps were derived from [^{15}N,^{1}H]-HSQC experiments[72] utilizing pulsed-field gradients for coherence selection and quadrature detection[73] and incorporating the sensitivity enhancement element of Rance and Palmer.[51,74] The ^{15}N{^{1}H}-NOE data were measured using a proton-detected version of the ^{15}N{^{1}H} steady-state heteronuclear Overhauser effect.[75] All experiments were recorded on a Bruker AV 700-MHz spectrometer equipped with a triple-resonance cryoprobe at 310 K. Spectra were processed and analyzed in the spectrometer software TOPSPIN 1.3 and calibrated relative to the water resonance at 4.63 ppm proton frequency, from which the ^{15}N scale was calculated indirectly.

ELISA

Biotinylated pD-peptide fusion proteins were immobilized on NeutrAvidin-coated plates after IMAC purification using 200 μl of 10-μM protein solutions and 1 h incubation time. One hundred microliters of 1 μM armadillo repeat proteins was incubated for 1 h. Binding was detected with an anti-$MRGSH_4$ antibody (Qiagen), a secondary anti-mouse immunoglobulin G alkaline phosphatase conjugate (Sigma), and *p*-nitrophenylphosphate (Fluka). Absorbance at 405 nm was measured using a Perkin Elmer HTS 7000 Plus plate reader. A buffer solution containing 50 mM Tris–HCl, 150 mM NaCl, and 0.5% bovine serum albumin (pH 7.4) was used for all the proteins and for the blocking steps. Washing after each step was carried out with $TBST_{150}$ (Tris–HCl 50 mM, NaCl 150 mM, and 0.05% Tween 20, pH 7.4). All steps were carried out at 4 °C. Development with 4-nitrophenylphosphate and readout were performed at room temperature.

Acknowledgements

The authors want to thank W.I. Weis, M. Köhler, and E. Conti for kindly providing the plasmids containing the natural armadillo repeat protein genes. We thank Dr. P. Kolb for valuable suggestions, Dr. A. Honegger for EXCEL macros, and the other members of the Plückthun laboratory for fruitful discussions. The calculations were performed on Matterhorn, a Beowulf Linux cluster at the Informatikdienste of the University of Zürich. We thank C. Bolliger, Dr. T. Steenbock, and Dr. A. Godknecht for installing and maintaining the Linux cluster. F. Parmeggiani was the recipient of a predoctoral fellowship from the Roche Research Foundation. F.P. and G.V. are members of the Molecular Life Science Ph.D. program. This work was supported by the Swiss National Center of Competence in Research (NCCR) in Structural Biology and in part by a Discovery grant from the Kommission für Technologie und Innovation (KTI).

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2007.12.014

References

1. Hoogenboom, H. R. (2005). Selecting and screening recombinant antibody libraries. *Nat. Biotechnol.* **23**, 1105–1116.
2. Binz, H. K., Amstutz, P. & Plückthun, A. (2005). Engineering novel binding proteins from nonimmunoglobulin domains. *Nat. Biotechnol.* **23**, 1257–1268.
3. Almagro, J. C. (2004). Identification of differences in the specificity-determining residues of antibodies that recognize antigens of different size: implications for the rational design of antibody repertoires. *J. Mol. Recognit.* **17**, 132–143.
4. MacCallum, R. M., Martin, A. C. & Thornton, J. M. (1996). Antibody–antigen interactions: contact analysis and binding site topography. *J. Mol. Biol.* **262**, 732–745.
5. Marchalonis, J. J., Adelman, M. K., Robey, I. F., Schluter, S. F. & Edmundson, A. B. (2001). Exquisite specificity and peptide epitope recognition promiscuity, properties shared by antibodies from sharks to humans. *J. Mol. Recognit.* **14**, 110–121.
6. Wilson, I. A., Ghiara, J. B. & Stanfield, R. L. (1994). Structure of anti-peptide antibody complexes. *Res. Immunol.* **145**, 73–78.
7. Kuriyan, J. & Cowburn, D. (1997). Modular peptide recognition domains in eukaryotic signaling. *Annu. Rev. Biophys. Biomol. Struct.* **26**, 259–288.
8. Esteban, Ö. & Zhao, H. (2004). Directed evolution of soluble single-chain human class II MHC molecules. *J. Mol. Biol.* **340**, 81–95.
9. Blatch, G. L. & Lässle, M. (1999). The tetratricopeptide repeat: a structural motif mediating protein–protein interactions. *BioEssays*, **21**, 932–939.
10. Coates, J. C. (2003). Armadillo repeat proteins: beyond the animal kingdom. *Trends Cell Biol.* **13**, 463–471.
11. Smith, T. F., Gaitatzes, C., Saxena, K. & Neer, E. J. (1999). The WD repeat: a common architecture for diverse functions. *Trends Biochem. Sci.* **24**, 181–185.
12. Peifer, M., Berg, S. & Reynolds, A. B. (1994). A repeating amino acid motif shared by proteins with diverse cellular roles. *Cell*, **76**, 789–791.
13. Hatzfeld, M. (1999). The armadillo family of structural proteins. *Int. Rev. Cytol.* **186**, 179–224.
14. Harris, T. J. & Peifer, M. (2005). Decisions, decisions: beta-catenin chooses between adhesion and transcription. *Trends Cell Biol.* **15**, 234–237.
15. Anastasiadis, P. Z. & Reynolds, A. B. (2000). The p120 catenin family: complex roles in adhesion, signaling and cancer. *J. Cell Sci.* **113**, 1319–1334.
16. Nathke, I. S. (2004). The adenomatous polyposis coli protein: the Achilles heel of the gut epithelium. *Annu. Rev. Cell Dev. Biol.* **20**, 337–366.
17. Goldfarb, D. S., Corbett, A. H., Mason, D. A., Harreman, M. T. & Adam, S. A. (2004). Importin alpha: a multipurpose nuclear-transport receptor. *Trends Cell Biol.* **14**, 505–514.
18. Wieschaus, E., Nüsslein-Volhard, C. & Jürgens, G. (1984). Mutations affecting the pattern of the larval cuticle in *Drosophila melanogaster*. 3. Zygotic loci on the X-chromosome and 4th chromosome. *Wilhelm Roux's Arch. Dev. Biol.* **193**, 296–307.
19. Riggleman, B., Wieschaus, E. & Schedl, P. (1989). Molecular analysis of the armadillo locus: uniformly distributed transcripts and a protein with novel internal repeats are associated with a *Drosophila* segment polarity gene. *Genes Dev.* **3**, 96–113.
20. Groves, M. R. & Barford, D. (1999). Topological characteristics of helical repeat proteins. *Curr. Opin. Struct. Biol.* **9**, 383–389.
21. Kobe, B. & Kajava, A. V. (2000). When protein folding is simplified to protein coiling: the continuum of solenoid protein structures. *Trends Biochem. Sci.* **25**, 509–515.
22. Huber, A. H., Nelson, W. J. & Weis, W. I. (1997). Three-dimensional structure of the armadillo repeat region of beta-catenin. *Cell*, **90**, 871–882.
23. Conti, E., Uy, M., Leighton, L., Blobel, G. & Kuriyan, J. (1998). Crystallographic analysis of the recognition of a nuclear localization signal by the nuclear import factor karyopherin alpha. *Cell*, **94**, 193–204.
24. Catimel, B., Teh, T., Fontes, M. R., Jennings, I. G., Jans, D. A., Howlett, G. J. *et al.* (2001). Biophysical characterization of interactions involving importin-alpha during nuclear import. *J. Biol. Chem.* **276**, 34189–34198.
25. Forrer, P., Stumpp, M. T., Binz, H. K. & Plückthun, A. (2003). A novel strategy to design binding molecules harnessing the modular nature of repeat proteins. *FEBS Lett.* **539**, 2–6.
26. Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R. D. & Bairoch, A. (2003). ExPASy: the proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res.* **31**, 3784–3788.
27. Binz, H. K., Stumpp, M. T., Forrer, P., Amstutz, P. & Plückthun, A. (2003). Designing repeat proteins: well-expressed, soluble and stable proteins from combinatorial libraries of consensus ankyrin repeat proteins. *J. Mol. Biol.* **332**, 489–503.
28. Mosavi, L. K., Minor, D. L., Jr & Peng, Z. Y. (2002). Consensus-derived structural determinants of the ankyrin repeat motif. *Proc. Natl Acad. Sci. USA*, **99**, 16029–16034.
29. Main, E. R., Xiong, Y., Cocco, M. J., D'Andrea, L. & Regan, L. (2003). Design of stable alpha-helical arrays from an idealized TPR motif. *Structure*, **11**, 497–508.

30. Stumpp, M. T., Forrer, P., Binz, H. K. & Plückthun, A. (2003). Designing repeat proteins: modular leucine-rich repeat protein libraries based on the mammalian ribonuclease inhibitor family. *J. Mol. Biol.* **332**, 471–487.
31. Interlandi, G., Wetzel, S. K., Settanni, G., Plückthun, A. & Caflisch, A. (2008). Characterization and further stabilization of designed ankyrin repeat proteins by combining molecular dynamics simulations and experiments. *J. Mol. Biol.* **375**, 837–854.
32. Andrade, M. A., Petosa, C., O'Donoghue, S. I., Müller, C. W. & Bork, P. (2001). Comparison of ARM and HEAT protein repeats. *J. Mol. Biol.* **309**, 1–18.
33. Schultz, J., Milpetz, F., Bork, P. & Ponting, C. P. (1998). SMART, a simple modular architecture research tool: identification of signaling domains. *Proc. Natl Acad. Sci. USA*, **95**, 5857–5864.
34. Letunic, I., Copley, R. R., Pils, B., Pinkert, S., Schultz, J. & Bork, P. (2006). SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res.* **34**, D257–D260.
35. Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680.
36. Lange, A., Mills, R. E., Lange, C. J., Stewart, M., Devine, S. E. & Corbett, A. H. (2007). Classical nuclear localization signals: definition, function, and interaction with importin alpha. *J. Biol. Chem.* **282**, 5101–5105.
37. Xu, W. & Kimelman, D. (2007). Mechanistic insights from structural studies of beta-catenin and its binding partners. *J. Cell Sci.* **120**, 3337–3344.
38. von Kries, J. P., Winbeck, G., Asbrand, C., Schwarz-Romond, T., Sochnikova, N., Dell'Oro, A. *et al.* (2000). Hot spots in beta-catenin for interactions with LEF-1, conductin and APC. *Nat. Struct. Biol.* **7**, 800–807.
39. Hoffmans, R. & Basler, K. (2004). Identification and *in vivo* role of the Armadillo–Legless interaction. *Development*, **131**, 4393–4400.
40. Leung, S. W., Harreman, M. T., Hodel, M. R., Hodel, A. E. & Corbett, A. H. (2003). Dissection of the karyopherin alpha nuclear localization signal (NLS)-binding groove: functional requirements for NLS binding. *J. Biol. Chem.* **278**, 41947–41953.
41. Sreerama, N. & Woody, R. W. (2004). Computation and analysis of protein circular dichroism spectra. *Methods Enzymol.* **383**, 318–351.
42. Slavik, J. (1982). Anilinonaphthalene sulfonate as a probe of membrane composition and function. *Biochim. Biophys. Acta*, **694**, 1–25.
43. Ptitsyn, O. B. (1995). Molten globule and protein folding. *Adv. Protein Chem.* **47**, 83–229.
44. Butterfoss, G. L. & Kuhlman, B. (2006). Computer-based design of novel protein structures. *Annu. Rev. Biophys. Biomol. Struct.* **35**, 49–65.
45. Voigt, C. A., Gordon, D. B. & Mayo, S. L. (2000). Trading accuracy for speed: a quantitative comparison of search algorithms in protein sequence design. *J. Mol. Biol.* **299**, 789–803.
46. Conti, E. & Kuriyan, J. (2000). Crystallographic analysis of the specific yet versatile recognition of distinct nuclear localization signals by karyopherin alpha. *Structure*, **8**, 329–338.
47. Fontes, M. R., Teh, T., Toth, G., John, A., Pavo, I., Jans, D. A. & Kobe, B. (2003). Role of flanking sequences and phosphorylation in the recognition of the simian-virus-40 large T-antigen nuclear localization sequences by importin-alpha. *Biochem. J.* **375**, 339–349.
48. Chothia, C. (1975). Structural invariants in protein folding. *Nature*, **254**, 304–308.
49. Baum, J., Dobson, C. M., Evans, P. A. & Hanley, C. (1989). Characterization of a partly folded protein by NMR methods—studies on the molten globule state of guinea-pig alpha-lactalbumin. *Biochemistry*, **28**, 7–13.
50. Dyson, H. J. & Wright, P. E. (1998). Equilibrium NMR studies of unfolded and partially folded proteins. *Nat. Struct. Biol.* **5**, 499–503.
51. Palmer, A. G. (2001). NMR probes of molecular dynamics: overview and comparison with other techniques. *Annu. Rev. Biophys. Biomol. Struct.* **30**, 129–155.
52. Kalderon, D., Roberts, B. L., Richardson, W. D. & Smith, A. E. (1984). A short amino acid sequence able to specify nuclear location. *Cell*, **39**, 499–509.
53. Choi, H. J. & Weis, W. I. (2005). Structure of the armadillo repeat domain of plakophilin 1. *J. Mol. Biol.* **346**, 367–376.
54. Mosavi, L. K. & Peng, Z. Y. (2003). Structure-based substitutions for increased solubility of a designed protein. *Protein Eng.* **16**, 739–745.
55. Main, E. R., Stott, K., Jackson, S. E. & Regan, L. (2005). Local and long-range stability in tandemly arrayed tetratricopeptide repeats. *Proc. Natl Acad. Sci. USA*, **102**, 5721–5726.
56. Kajander, T., Cortajarena, A. L., Main, E. R., Mochrie, S. G. & Regan, L. (2005). A new folding paradigm for repeat proteins. *J. Am. Chem. Soc.* **127**, 10188–10190.
57. Wetzel, S. K., Settanni, G., Kenig, M., Binz, H. K. & Plückthun, A. (2008). Folding and unfolding mechanism of highly stable full consensus ankyrin repeat proteins. *J. Mol. Biol.* In press. doi:10.1016/j.jmb.2007.11.046
58. Eriksson, A. E., Baase, W. A., Zhang, X. J., Heinz, D. W., Blaber, M., Baldwin, E. P. & Matthews, B. W. (1992). Response of a protein structure to cavity-creating mutations and its relation to the hydrophobic effect. *Science*, **255**, 178–183.
59. Schueler-Furman, O., Wang, C., Bradley, P., Misura, K. & Baker, D. (2005). Progress in modeling of protein structures and interactions. *Science*, **310**, 638–642.
60. Desjarlais, J. R. & Handel, T. M. (1999). Side-chain and backbone flexibility in protein core design. *J. Mol. Biol.* **290**, 305–318.
61. Looger, L. L., Dwyer, M. A., Smith, J. J. & Hellinga, H. W. (2003). Computational design of receptor and sensor proteins with novel functions. *Nature*, **423**, 185–190.
62. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H. *et al.* (2000). The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242.
63. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.
64. McGinnis, S. & Madden, T. L. (2004). BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res.* **32**, W20–W25.
65. Guex, N. & Peitsch, M. C. (1997). SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis*, **18**, 2714–2723.
66. Koradi, R., Billeter, M. & Wüthrich, K. (1996). MOLMOL: a program for display and analysis of macromolecular structures. *J. Mol. Graphics*, **14**, 51–55, 29–32.
67. Sambrook, J. & Russell, D. W. (2001). *Molecular Cloning: A Laboratory Manual*, 3rd edit., Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
68. Inoue, H., Nojima, H. & Okayama, H. (1990). High efficiency transformation of *Escherichia coli* with plasmids. *Gene*, **96**, 23–28.

69. Köhler, M., Speck, C., Christiansen, M., Bischoff, F. R., Prehn, S., Haller, H. *et al.* (1999). Evidence for distinct substrate specificities of importin alpha family members in nuclear protein import. *Mol. Cell. Biol.* **19**, 7782–7791.
70. Cull, M. G. & Schatz, P. J. (2000). Biotinylation of proteins *in vivo* and *in vitro* using small peptide tags. *Methods Enzymol.* **326**, 430–440.
71. Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. & Karplus, M. (1983). CHARMM—a program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **4**, 187–217.
72. Bodenhausen, G. & Ruben, D. J. (1980). Natural abundance N-15 NMR by enhanced heteronuclear spectroscopy. *Chem. Phys. Lett.* **69**, 185–189.
73. Keeler, J., Clowes, R. T., Davis, A. L. & Laue, E. D. (1994). Pulsed-field gradients: theory and practice. *Methods Enzymol.* **239**, 145–207.
74. Kay, L. E., Keifer, P. & Saarien, T. (1992). Pure absorption gradient enhanced heteronuclear single-quantum correlation spectroscopy with improved sensitivity. *J. Am. Chem. Soc.* **114**, 10663–10665.
75. Noggle, J. H. & Schirmer, R. E. (1971). *The Nuclear Overhauser Effect: Chemical Applications,* Academic Press, New York.

Supplementary Materials

Designed Armadillo Repeat Proteins as General Peptide Binding Scaffolds: Consensus Design and Computational Optimization of the Hydrophobic Core

Fabio Parmeggiani [1], Riccardo Pellarin [1], Anders Peter Larsen [1], Gautham Varadamsetty [1], Michael T. Stumpp [1], Oliver Zerbe [2], Amedeo Caflisch [1] and Andreas Plückthun [1]*

[1] Department of Biochemistry, University of Zürich, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland

[2] Department of Organic Chemistry, University of Zürich, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland

* corresponding author

Present addresses: A. P. Larsen, Department of Biomedical Sciences, University of Copenhagen, Blegdamsvej 3, DK-2200 Copenhagen, Denmark; M. T. Stumpp, Molecular Partners AG, Grabenstrasse 11a, CH-8952 Schlieren, Switzerland

Email address of corresponding author: plueckthun@bioc.uzh.ch

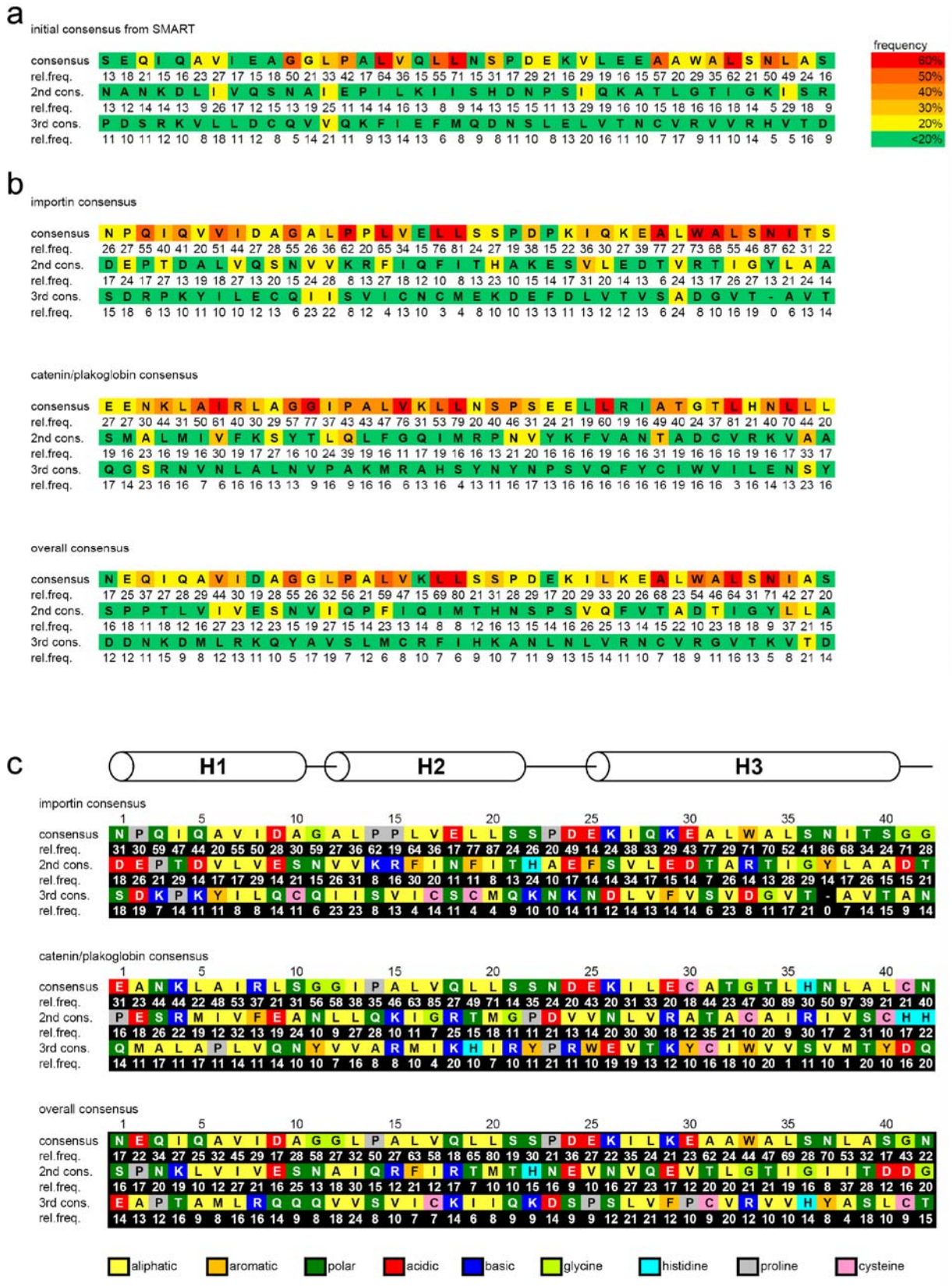
Residue Choices in Internal Repeats

Several residues present in the originally defined consensus sequences (Fig. S1) were replaced and some maintained despite their apparent unfavorable properties. The rationale behind these choices is illustrated here. The residues obtained from the consensus sequences were maintained in the positions not mentioned here.

Gln5, present in the consensus type I and C, can potentially form a hydrogen bond with Asp9 of the same repeat, stabilizing H1 (observed in structure 1EE5 [1]). Gly11 is conserved for its crucial role in bending the polypeptide chain between H1 and H2, being compatible with a positive φ angle required at this position. Pro14 is an unusual conserved feature of armadillo repeats, present at a frequency of 62% in the alignment of importin repeats, 35% for catenin/plakoglobin repeats and 50% in the overall alignment. It is located at the beginning of H2, at a position where it is still not necessary to involve the backbone nitrogen in a hydrogen bond; instead of disrupting the secondary structure, it adopts the φ/ψ angles typical for α-helices in the available crystal structures. Asn37 is a well conserved residue in all consensus sequences, due to its critical role in binding to the backbone of target peptides. Trp33 is also involved in binding, specifically in the recognition of target side chains in the case of importin-α proteins. Thus, it appears with high frequency in the importin subfamily and it is therefore present in the overall consensus as well. All the residues described above were maintained.

Further modifications were introduced in the original consensus sequences (Fig S1) to meet the requirements for protein production (e.g. lack of cysteines) or to avoid structural defects (e.g. presence of potential clashes) that could have arisen from a purely sequence-based alignment. Cys30 (18%) and Cys41 (21%) in the catenin consensus were replaced by the second most common amino acid (Ala 12% and His 17%, respectively), to avoid the formation of undesired disulfide bonds that might limit possible future applications.

Fig. S1 Consensus sequences derived from multiple alignments. (a) Consensus sequence obtained from alignment of SMART armadillo seed sequences. (b) Consensus sequences of importin and catenin/plakoglobin subfamilies and overall consensus; sequences used for the alignment were retrieved using a profile based on SMART sequences. The sequences are limited to 40 residues and do not contain the loop between adjacent repeats. Amino acids are colored according to their relative frequency. (c) Consensus sequences of importin and catenin/plakoglobin subfamilies and overall consensus after manual refinement of the alignment. The cylinders indicate the putative α-helices and the numbers denote the positions inside the single repeats according to the conventions introduced. The residues are colored according to amino acid type, as indicated at the bottom. For each position of the sequences, the most frequent, the second and the third most frequent type of residue are indicated, with the relative frequency expressed as percentage.


a
initial consensus from SMART
frequency
consensus S E Q I Q A V I E A G G L P A L V Q L L N S P D E K V L E E A A W A L S N L A S
rel.freq. 13 18 21 15 16 23 27 17 15 18 50 21 33 42 17 64 36 15 55 71 15 31 17 29 21 16 29 19 16 15 57 20 29 35 62 21 50 49 24 16
2nd cons. N A N K D L I V Q S N A I E P I L K I I S H D N P S I Q K A T L G T I G K I S R
rel.freq. 13 12 14 14 13 9 26 17 12 15 13 19 25 11 14 14 16 13 8 9 14 13 15 15 11 13 29 19 16 10 15 18 16 16 18 14 5 29 18 9
3rd cons. P D S R K V L L D C Q V V Q K F I E F M Q D N S L E L V T N C V R V V R H V T D
rel.freq. 11 10 11 12 10 8 18 11 12 8 5 14 21 11 9 13 14 13 6 8 9 8 11 10 8 13 20 16 11 10 7 17 9 11 10 14 5 5 16 9
60%
50%
40%
30%
20%
<20%
b
importin consensus
consensus N P Q I Q V V I D A G A L P P L V E L L S S P D P K I Q K E A L W A L S N I T S
rel.freq. 26 27 55 40 41 20 51 44 27 28 55 26 36 62 20 65 34 15 76 81 24 27 19 38 15 22 36 30 27 39 77 27 73 68 55 46 87 62 31 22
2nd cons. D E P T D A L V Q S N V V K R F I Q F I T H A K E S V L E D T V R T I G Y L A A
rel.freq. 17 24 17 27 13 19 18 27 13 20 15 24 28 8 13 27 18 12 10 8 13 23 10 15 14 17 31 20 14 13 6 24 13 17 26 27 13 21 24 14
3rd cons. S D R P K Y I L E C Q I I S V I C N C M E K D E F D L V T V S A D G V T - A V T
rel.freq. 15 18 6 13 10 11 10 10 12 13 6 23 22 8 12 4 13 10 3 4 8 10 10 13 13 11 13 12 12 13 6 24 8 10 16 19 0 6 13 14
catenin/plakoglobin consensus
consensus E E N K L A I R L A G G I P A L V K L L N S P S E E L L R I A T G T L H N L L L
rel.freq. 27 27 30 44 31 50 61 40 30 29 57 77 37 43 43 47 76 31 53 79 20 40 46 31 24 21 19 60 19 16 49 40 24 37 81 21 40 70 44 20
2nd cons. S M A L M I V F K S Y T L Q L F G Q I M R P N V Y K F V A N T A D C V R K V A A
rel.freq. 19 16 23 16 19 16 30 19 17 27 16 10 24 39 19 16 11 17 19 16 16 13 21 20 16 16 16 19 16 16 31 19 16 16 16 19 16 17 33 17
3rd cons. Q G S R N V N L A L N V P A K M R A H S Y N Y N P S V Q F Y C I W V I L E N S Y
rel.freq. 17 14 23 16 16 7 6 16 16 13 13 9 16 9 16 16 6 13 16 4 13 11 16 17 13 16 16 16 16 16 16 19 16 16 3 16 14 13 23 16
overall consensus
consensus N E Q I Q A V I D A G G L P A L V K L L S S P D E K I L K E A L W A L S N I A S
rel.freq. 17 25 37 27 28 29 44 30 19 28 55 26 32 56 21 59 47 15 69 80 21 31 28 29 17 20 29 33 20 26 68 23 54 46 64 31 71 42 27 20
2nd cons. S P P T L V I V E S N V I Q P F I Q I M T H N S P S V Q F V T A D T I G Y L L A
rel.freq. 16 18 11 18 12 16 27 23 12 23 15 19 27 15 14 23 13 14 8 8 12 16 13 15 14 16 26 25 13 14 15 22 10 23 18 18 9 37 21 15
3rd cons. D D N K D M L R K Q Y A V S L M C R F I H K A N L N L V R N C V R G V T K V T D
rel.freq. 12 12 11 15 9 8 12 13 11 10 5 17 19 7 12 6 8 10 7 6 9 10 7 11 9 13 15 14 11 10 7 18 9 11 16 13 5 8 21 14
c
H1
H2
H3
importin consensus
1 5 10 15 20 25 30 35 40
consensus N P Q I Q A V I D A G A L P P L V E L L S S P D E K I Q K E A L W A L S N I T S G G
rel.freq. 31 30 59 47 44 20 55 50 28 30 59 27 36 62 19 64 36 17 77 87 24 26 20 49 14 24 38 33 29 43 77 29 71 70 52 41 86 68 34 24 71 28
2nd cons. D E P T D V L V E S N V V K R F I N F I T H A E F S V L E D T A R T I G Y L A A D T
rel.freq. 18 26 21 29 14 17 17 29 14 21 15 26 31 8 16 30 20 11 11 8 13 24 10 17 14 14 34 17 15 14 7 26 14 13 28 29 14 17 26 15 15 21
3rd cons. S D K P K Y I L Q C Q I I S V I C S C M Q K N K N D L V F V S V D G V T - A V T A N
rel.freq. 18 19 7 14 11 11 8 8 14 11 6 23 23 8 13 4 14 11 4 4 9 10 10 14 11 12 14 13 14 14 6 23 8 11 17 21 0 7 14 15 9 14
catenin/plakoglobin consensus
1 5 10 15 20 25 30 35 40
consensus E A N K L A I R L S G G I P A L V Q L L S S N D E K I L E C A T G T L H N L A L C N
rel.freq. 31 23 44 44 22 48 53 37 21 31 56 58 38 35 46 63 85 27 49 71 14 35 24 20 43 20 31 33 20 18 44 23 47 30 89 30 50 97 39 21 21 40
2nd cons. P E S R M I V F E A N L L Q K I G R T M G P D V V N L V R A T A C A I R I V S C H H
rel.freq. 16 18 26 22 19 12 32 13 19 24 10 9 27 28 10 11 7 25 15 18 11 11 21 13 14 20 30 30 18 12 35 21 10 20 9 30 17 2 31 10 17 22
3rd cons. Q M A L A P L V Q N Y V V A R M I K H I R Y P R W E V T K Y C I W V V S V M T Y D Q
rel.freq. 14 11 17 11 17 11 14 11 14 10 10 7 16 8 8 10 4 20 10 7 10 11 21 11 10 19 19 13 12 10 16 18 10 20 1 11 10 1 20 10 16 20
overall consensus
1 5 10 15 20 25 30 35 40
consensus N E Q I Q A V I D A G G L P A L V Q L L S S P D E K I L K E A A W A L S N L A S G N
rel.freq. 17 22 34 27 25 32 45 29 17 28 58 27 32 50 27 63 58 18 65 80 19 30 21 36 27 22 35 24 21 23 62 24 44 47 69 28 70 53 32 17 43 22
2nd cons. S P N K L V I V E S N A I Q R F I R T M T H N E V N V Q E V T L G T I G I I T D D G
rel.freq. 16 17 20 19 10 12 27 21 16 25 13 18 30 15 12 21 12 17 7 10 10 15 16 9 10 16 27 23 17 12 20 20 21 21 19 16 8 37 28 12 16 20
3rd cons. E A P T A M L R Q Q Q V V S V I C K I I Q K D S P S L V F P C V R V V H Y A S L C T
rel.freq. 14 13 12 16 9 8 16 16 14 9 8 18 24 8 10 7 7 14 6 8 9 9 14 9 9 12 21 21 12 10 9 20 12 10 10 14 8 4 18 10 9 15
aliphatic
aromatic
polar
acidic
basic
glycine
histidine
proline
cysteine

Pro2 (30%), in the importin consensus, was substituted with the approximately equally conserved Glu (26%), as Pro at such position would probably disrupt H1, as indicated by importin-α crystal structures. In the catenin consensus, position 9 shows a preference for long aliphatic side-chains, either non-polar or polar (Leu, Glu, Gln); however, this residue is solvent-exposed, and Leu9 (21%) was substituted by the second most common amino acid Glu (19%). In the importin consensus, Pro15 (19%) (Fig. S1) was derived from sequences which do not possess a Pro in position 14. A double Pro14-Pro15 never occurs in the observed sequences and it is likely to be extremely destabilizing for H2. Position 15 is usually occupied by small hydrophobic residues in combination with Pro14. Arg also represents a relatively common choice (16%) but it occurs almost exclusively in the second repeat of natural importins. Val was therefore chosen as more general substitution, instead of Arg, due to the slightly higher frequency of occurrence (13%) compared to other residues. The catenin consensus has Gln as most frequent amino acid (27%) at position 18. However, both Arg and Lys are represented almost at the same frequency (25% and 20% respectively), indicating a preference for positively charged residues. Arg was thus chosen at this position due to its higher frequency. Positions 24 and 25 at the joint between loop H2-H3 and H3 show a clear preference for acidic residues in all consensus sequences. However, a pair of negatively charged residues never occurs in the observed sequences, and it could lead to charge repulsion or formation of a negatively charged belt along the whole protein. The most conserved residues in importin and catenin consensus sequences were preserved (Asp24 in the importin consensus (49%) and Glu25 in the catenin/plakoglobin consensus (43%), respectively). To reduce the local negative charge, Asn (11%) was chosen to replace Glu (19%) at position 25 in the importin consensus, because it is the second most frequent polar residue. Asp24 in the catenin/plakoglobin consensus (20%) was substituted by Asn (10%), a choice driven by the similarity to the original residue in a pool of candidates with almost the same frequency (Arg, Asn, Met, Ser, Tyr, Val). In the overall consensus, a preferred position for the negative charge is not pronounced (Asp24 36%, Glu25 27% in type C), and the alternative amino acids have all very low frequency (<10%). The residues selected for replacement were thus chosen to improve the H3 stability: Glu25 was kept in the overall consensus due to its higher helical propensity and Asn24 was introduced to keep structural similarity to the more frequent residue Asp and to take advantage of its propensity as an N-cap residue [2-5]. Gly was introduced at position 42 for cloning purposes. It is noteworthy that, because of short H3-H1 loops (1-3 residues in general), one position of the loop is very often occupied by a Gly. Taking into

account that position 41 is sometimes involved in binding and will then possibly be subjected to mutation for applications, it is important to keep a constant glycine inside the loop to maintain the required flexibility.

Designed capping repeats

Capping repeats were designed based on type C internal repeat. A detailed description of the residues introduced in the capping repeats is provided here. The original residues of the type C repeat were maintained in the positions which are not described.

The N-terminal designed capping repeat (Na) goes from position 12 to 42 and includes only H2 and H3. Positions 12, 19, 27, 34 are occupied by hydrophobic residues in the consensus and had to be replaced by hydrophilic residues based on structures and common residues obtained from alignment of N-terminal capping sequences. Ser12 provides the N-terminal helix cap of H2. Asn14 substitutes the more common proline, providing a polar residue with a relatively short side chain. Glu15 can interact with Ser12 in a helix and can additionally stabilize it [6]. Lys18 can form a salt bridge with Glu15, stabilizing the helix, and, in general, a long polar residue is required at this position. Gln19 provides a hydrophobic part for interaction with the neighboring internal repeat as well as a polar moiety for solvent exposure. Asn21 is common at this position and it has a good propensity as helix C-capping residue. Asp23 and Asp24 are conserved as a charged couple in several N-terminal capping repeats. Gln25 is well conserved, polar and with high helical propensity. Gln27 provides a hydrophobic part for interaction with the neighboring internal repeat as well as a polar moiety for solvent exposure. Gln33, well conserved among capping repeats, substitutes the Trp, present at high frequency in internal modules because involved in binding in importins. Lys34 is present at moderate frequency in N-terminal capping repeats, among other polar residues. Arg36 has a high frequency of occurrence and seems to be able to interact with Trp33 present in the importin and overall consensus. Gln37 has a long side chain typical for residues at this position, and, instead of the more common lysine, avoids the formation of a positively charged spot in combination with Arg36. Asp41 maintains the negative charge often present in this position and breaks the helix. Gly42 was introduced to add flexibility and for further module assembly.

The C-terminal designed capping repeat (Ca) includes all three helices. Positions 8, 13, 17, 20, 28, 32, 35, 38, 39 are occupied by hydrophobic residues in the consensus and had to be replaced by hydrophilic residues based on structures and common residues obtained from

alignment of C-terminal capping sequences. Lys4 can potentially contribute to the hydrophobic core with the long aliphatic part of the side chain, while contacting the solvent with the positively charged amino-group. Lys8 is present to avoid the formation of a cluster of negative charges that would be formed if a conserved glutamate was used at this position, while keeping a high helical propensity and a long side chain. Glu9 is a highly conserved residue. Ala12 can potentially interact with the hydrophobic core. Glu14 is a common polar residue with high helical propensity as substitute for proline. Lys15 is also a common polar residue with relatively high helical propensity. Leu13 occupies a former hydrophobic core position, but it was retained for its high helical propensity and its ability to interact with Phe39. Glu17 corresponds to a core position in an internal repeat; the hydrophobic residue was substituted with this frequently occurring hydrophilic amino acid with high helical propensity. Gln20 is the more frequent polar residue used to substitute the conserved leucine present in the internal repeats. Positions from 21 to 23 are not clearly defined, showing strong conservation in the catenin/plakoglobin subfamily (maybe for functional reasons) and higher degree of variability in importins. The most conserved residues from the importin subfamily were thus chosen to occupy these positions. Gln28 can provide hydrophobic interactions and a polar side chain, and represents a better choice compared to a conserved tyrosine in the importin subfamily and an alanine in the catenin/plakoglobin subfamily. Gln32 provides high helical propensity and a polar side chain. The presence of frequent aromatic residues at this position does not seem to have a structural reason, judging from the crystal structures. Glu33 is one of the charged residues often found at this position and it has high helical propensity. Glu36 has high frequency in importins where this position is occupied by acidic residues, while in the catenin/plakoglobin subfamily phenylalanine and tyrosine are present. The aromatic residues have probably a functional role, but in our case a charged residue constitutes the better choice, because of the exposed position. Lys37 was chosen to replace the conserved asparagine in the internal repeats. Gln38 was chosen to replace hydrophobic residues, always present at this position both in internal and capping repeats, while providing a polar moiety in contact with the solvent. Phe39 is conserved in several capping repeats. From the available structures, it seems to be important for sealing of the hydrophobic core of importins and for compactness of C-terminal capping repeat via interaction with Leu13. His41 has been added as capping residue to stabilize H3.

References

1. Conti, E. & Kuriyan, J. (2000). Crystallographic analysis of the specific yet versatile recognition of distinct nuclear localization signals by karyopherin alpha. *Structure Fold Des* **8**, 329-38.
2. Serrano, L., Sancho, J., Hirshberg, M. & Fersht, A. R. (1992). Alpha-helix stability in proteins. I. Empirical correlations concerning substitution of side-chains at the N and C-caps and the replacement of alanine by glycine or serine at solvent-exposed surfaces. *J Mol Biol* **227**, 544-59.
3. Penel, S., Hughes, E. & Doig, A. J. (1999). Side-chain structures in the first turn of the alpha-helix. *J Mol Biol* **287**, 127-43.
4. Doig, A. J. & Baldwin, R. L. (1995). N- and C-capping preferences for all 20 amino acids in alpha-helical peptides. *Protein Sci* **4**, 1325-36.
5. Richardson, J. S. & Richardson, D. C. (1988). Amino acid preferences for specific locations at the ends of alpha helices. *Science* **240**, 1648-52.
6. Harper, E. T. & Rose, G. D. (1993). Helix stop signals in proteins and peptides: the capping box. *Biochemistry* **32**, 7605-9.

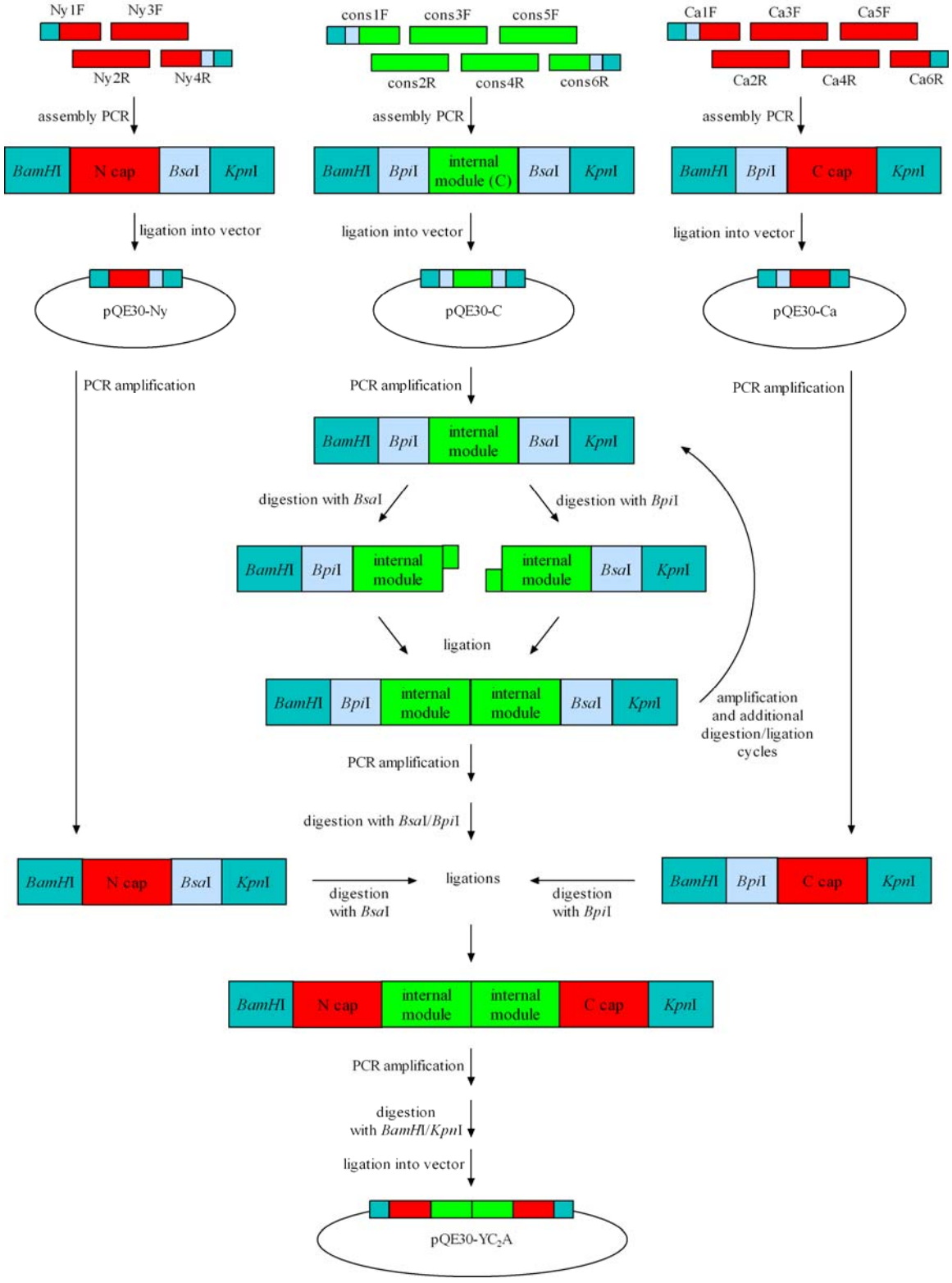


Fig. S2 Scheme of the assembly strategy for designed armadillo repeat protein constructs at the DNA level. Oligonucleotides are assembled to an internal or terminal capping module by PCR. The single modules contain external restriction sites for *BamH*I and *Kpn*I for insertion in the vector and sites for the type IIS restriction enzymes *Bsa*I and *Bpi*I for ligation of the modules. The construction of YC_2A from the internal module C, the N-terminal cap Ny and the C-terminal cap Ca is shown as an example.

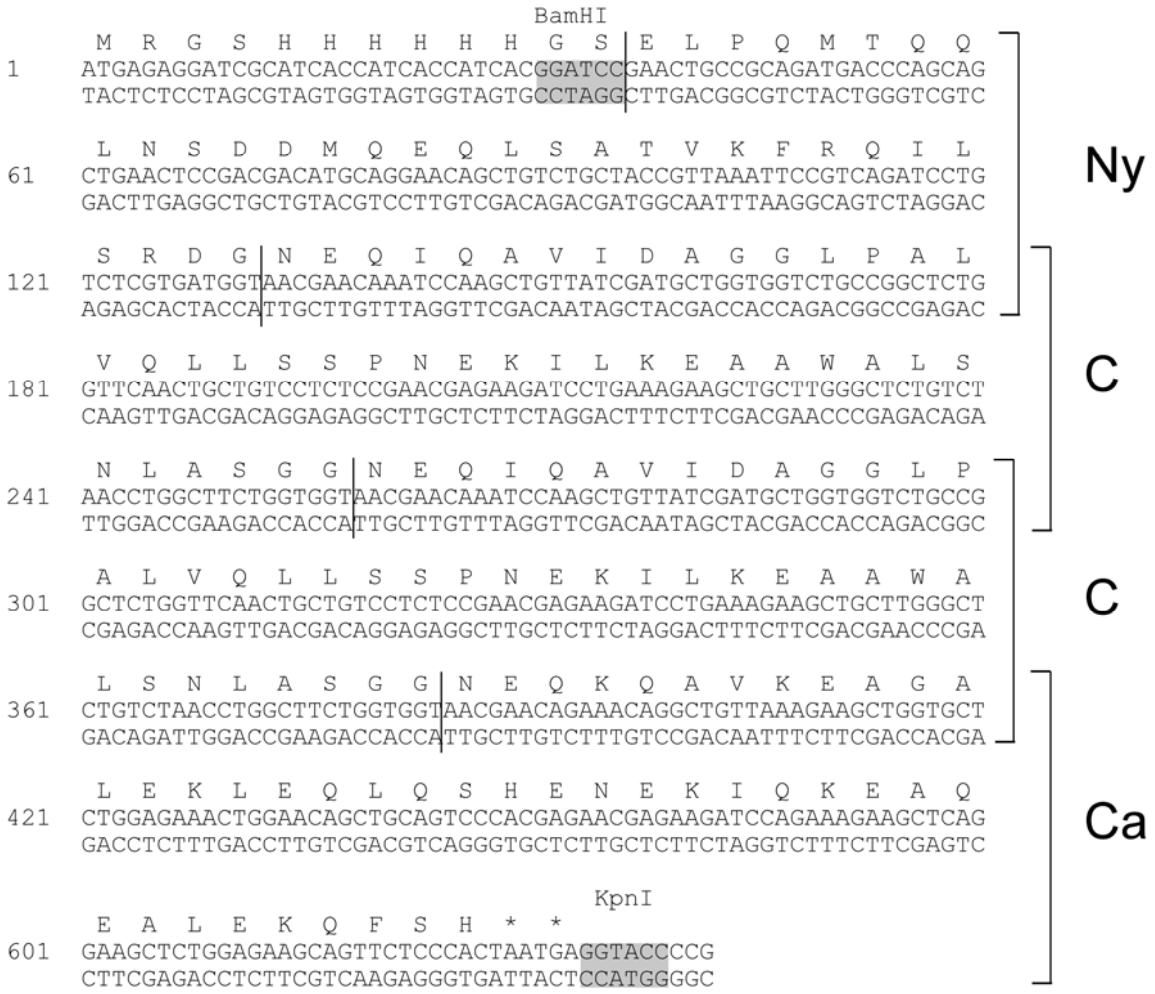


Fig. S3 Sequence of the designed armadillo repeat protein YC_2A. The translated amino acid sequence is shown on the top of the DNA sequence as single letter code. The bars indicate the separations between the modules and between the $MRGSHis_6$ tag provided by the vector and the N-terminal capping module. The modules are indicated on the right. The restriction sites used for cloning are highlighted in gray. The star (*) indicates the presence of a stop codon.

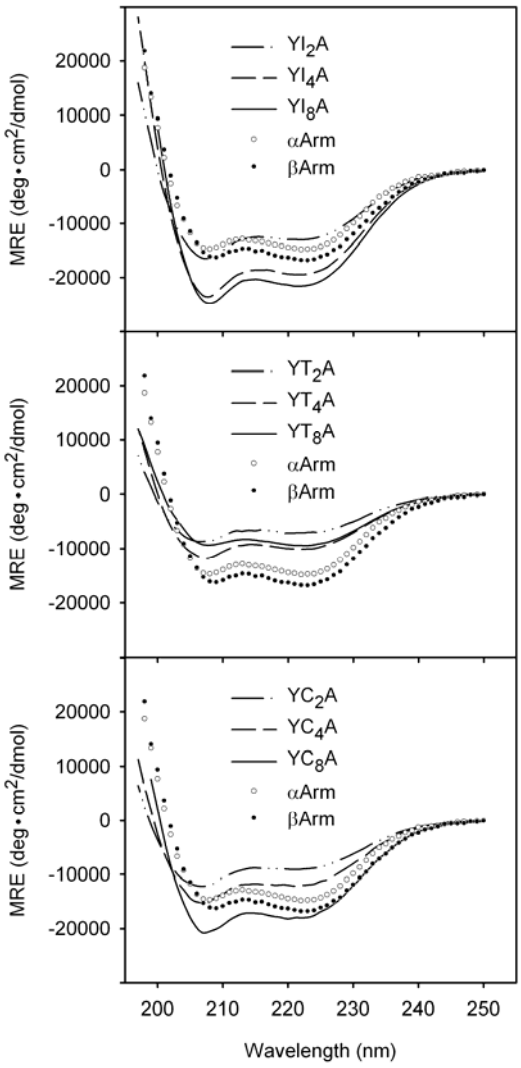


Fig. S4 Circular dichroism (CD) spectra of designed consensus armadillo repeat proteins. From the top, I-type, T-type and C-type proteins containing 2, 4 or 8 internal modules are shown. The CD spectra of the natural armadillo domains of human importin-α1 (α Arm) and mouse β-catenin (βArm) are indicated by empty and filled circles, respectively. The values are reported as mean residue ellipticity (MRE).

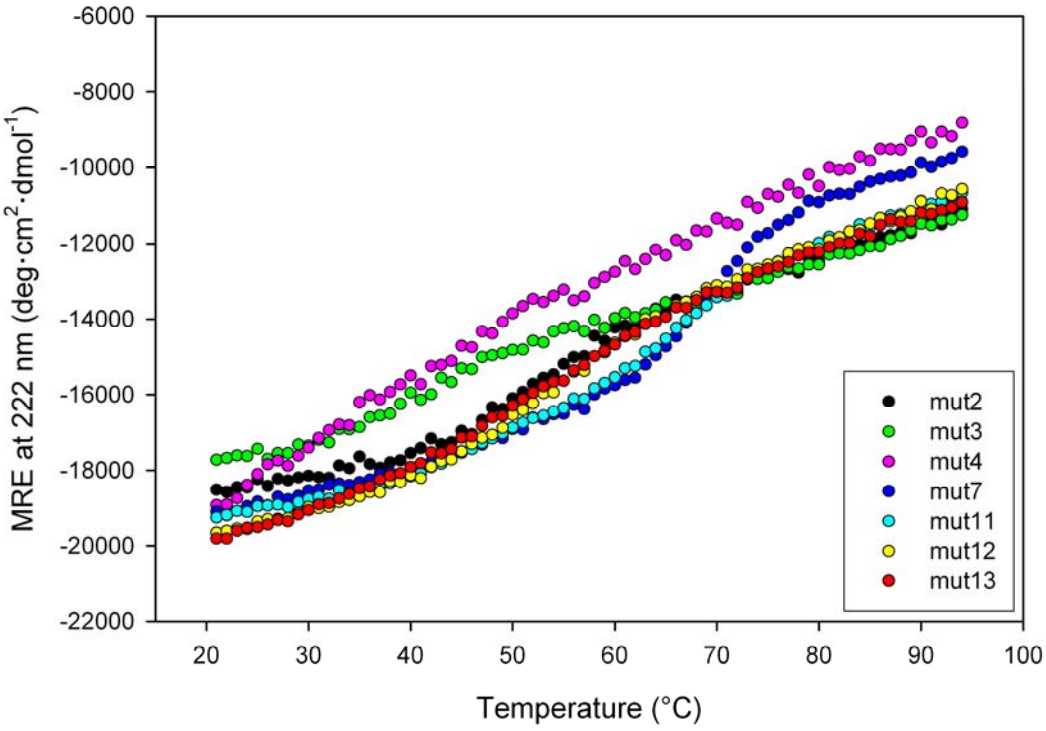


Fig. S5 Thermal denaturation of hydrophobic core mutants. The most promising mutants, based on size exclusion chromatography and ANS binding experiments, are shown. The CD signal at 222 nm is reported as mean residue ellipticity (MRE) as a function of temperature. Remarkably, all the mutants have similar ellipticity at 20°C. mut7 is characterized by the steepest and largest transition.

Table S1: Oligonucleotides used for the assembly and cloning of designed and natural armadillo repeat protein genes

name	sequence 5'-3' direction	description (for=forward, rev=reverse)
AcatFOR	CGGGATCCACACGTGCAATTCCTG	for β-catenin mouse
AcatREV	GCGGTACCATTAGTCCTCAGACATTCGG	rev β-catenin mouse
IMAF5	CGGGATCCCATCACTTCTGACATGATTGAG	for importin-α1 human
IMAR5	GCGGTACCATTACCCGAAGTAATGCTCAATAAG	rev importin-α1 human
pQE_f_1	CGGATAACAATTTCACACAG	forward primer for pQE vectors
pQE_r_1	GTTCTGAGGTCATTACTG	reverse primer for pQE vectors
Ny1F	CCAGGGATCCGAACTGCCGCAGATGACCCAGCAGCTGAACTCTG	for assembly Ny module and amplification
Ny2R	CGGTAGCAGACAGCTGTTCCTGCATGTCGTCAGAGTTCAGCTGCTGGG	rev assembly Ny module
Ny3F	GAACAGCTGTCTGCTACCGTTAAATTCCGTCAGATCCTGTCTCGTGATGG	for assembly Ny module
Ny4R	TTCCTGGTACCCTAAGGTCTCAACCATCACGAGACAGGATCTG	rev assembly Ny module and amplification
Na1F	CCAGGGATCCTCTCTGAACGAACTGGTTAAACAGCTGAACTCCG	for assembly Na module and amplification
Na2R	CTGAGCAGCTTCTTTCAGCTGTTTCTGGTCGTCGGAGTTCAGCTGTTTAACCAG	rev assembly Na module
Na3F	CAGCTGAAAGAAGCTGAAAGAAGCTGCTCAGAAACTGCGTCAGCTGGCTTCCGATGG	for assembly Na module
Na4R	TTCCTGGTACCCTAAGGTCTCAACCATCGGAAGCCAGCTG	rev assembly Na module and amplification
Cy1F	CCAGGGATCCTAGGAAGACCTTGGTGACAACATCAACG	for assembly Cy module and amplification
Cy2R	GCCACCAGCCTTCTCGATGAAGTCCGCGTTCTCGTTGATGTTGTCACCAAGG	rev assembly Cy module
Cy3F	CGAGAAGGCTGGTGGCATGGAGAAGATCTTCAACTGCCAGCAGAACG	for assembly Cy module
Cy4R	GCTTTCTCGTAGATCTTGTCGTTCTCGTTCTGCTGGCAGTTG	rev assembly Cy module
Cy5F	CGACAAGATCTACGAGAAAGCTTACAAGATCATCGAAACCTACTTCGGC	for assembly Cy module
Cy6R	TTCCTGGTACCTCATTAGCCGAAGTAGGTTTCGATG	rev assembly Cy module and amplification
CyM1F	CCAGGGATCCTAGGAAGACCTTGGTAACGAGAACGCGG	for assembly Cm module and amplification
CyM2R	GCCACCAGCCTTCTCGATGAAGTCCGCGTTCTCGTTACCAAGG	rev assembly Cm module
CyM3F	CGAGAAGGCTGGTGGCATGGAGAAGATCTTCAACGCTCAGCAGAACG	for assembly Cm module
CyM4R	GCTTTCTCGTAGATCTTGTCGTTCTCGTTCTGCTGAGCGTTG	rev assembly Cm module
Ca1F	CCAGGGATCCTAGGAAGACCTTGGTAACGAACAGAAACAGGC	for assembly Ca module and amplification

Table S1: Oligonucleotides used for the assembly and cloning of designed and natural armadillo repeat protein genes (continued)

name	sequence 5'-3' direction	description (for=forward, rev=reverse)
Ca2R	GTTTCTCCAGAGCACCAGCTTCTTTAACA GCCTGTTTCTGTTCGTTACC	rev assembly Ca module
Ca3F	GCTGGTGCTCTGGAGAAACTGGAACAGCT GCAGTCCCACGAG	for assembly Ca module
Ca4R	CCTGAGCTTCTTTCTGGATCTTCTCGTTC TCGTGGGACTGCAGC	rev assembly Ca module
Ca5F	GATCCAGAAAGAAGCTCAGGAAGCTCTGG AGAAGCAGTTCTCCC	for assembly Ca module
Ca6R	TTCCTGGTACCTCATTAGTGGGAGAACTG CTTCTCCAG	rev assembly Ca module and amplification
imp1F	CCAGGGATCCTAGGAAGACCTTGGTAACG AACAGATCC	for assembly importin module and amplification
imp2R	ACCGGCAGAGCACCAGCGTCGATAACAGC CTGGATCTGTTCGTTACCAAGG	rev assembly importin module
imp3F	CTGGTGCTCTGCCGGTTCTGGTTGAACTG CTGTCCTCTCCGGAC	for assembly importin module
imp4R	CCACAGAGCTTCTTTCTGGATCTTGTTGT CCGGAGAGGACAGCAG	rev assembly importin module
imp5F	TCCAGAAAGAAGCTCTGTGGGCTCTGTCT AACATCACTTCTGGTGGTTGAGACC	for assembly importin module
imp6R	TTCCTGGTACCCTAAGGTCTCAACCACCA GAAGTG	rev assembly importin module and amplification
cat1F	CCAGGGATCCTAGGAAGACCTTGGTGAAG C	for assembly catenin module and amplification
cat2R	CCACCAGATTCACGGATAGCCAGTTTGTT AGCTTCACCAAGGTCTTCC	rev assembly catenin module
cat3F	CTATCCGTGAATCTGGTGGTATCCCGGCT CTGGTTCGTCTGCTGTCCTC	for assembly catenin module
cat4R	TAGCAGCTTCCAGGATCTTCTCGTTGTTA GAGGACAGCAGACGAACC	rev assembly catenin module
cat5F	AAGATCCTGGAAGCTGCTACTGGCACTCT GCACAACCTGGCTCTGCATGGTTGAG	for assembly catenin module
cat6R	TTCCTGGTACCCTAAGGTCTCAACCATGC AGAGCC	rev assembly catenin module and amplification
cons1F	CCAGGGATCCTAGGAAGACCTTGGTAACG AACAAATCC	for assembly consensus module and amplification
cons2R	AGCCGGCAGACCACCAGCATCGATAACAG CTTGGATTTGTTCGTTACCAAGG	rev assembly consensus module
cons3F	GGTGGTCTGCCGGCTCTGGTTCAACTGCT GTCCTCTCCGAACG	for assembly consensus module
cons4R	CCAAGCAGCTTCTTTCAGGATCTTCTCGT TCGGAGAGGACAGC	rev assembly consensus module
cons5F	CCTGAAAGAAGCTGCTTGGGCTCTGTCTA ACCTGGCTTCTGGTGGTTGAG	for assembly consensus module
cons6R	TTCCTGGTACCCTAAGGTCTCAACCACCA GAAGCCAG	rev assembly consensus module and amplification
2A-rev	AGCCGGCAGAGCACCAGCATCGATAACAG CTTGGATTTGTTCGTTACCAAGG	rev hydrophobic core mutants assembly
2AVrev	AGCCGGAACAGCACCAGCATCGATAACAG CTTGGATTTGTTCGTTACCAAGG	rev hydrophobic core mutants assembly
2AIrev	AGCCGGGATAGCACCAGCATCGATAACAG CTTGGATTTGTTCGTTACCAAGG	rev hydrophobic core mutants assembly

Table S1: Oligonucleotides used for the assembly and cloning of designed and natural armadillo repeat protein genes (continued)

name	sequence 5'-3' direction	description (for=forward, rev=reverse)
3A-for	GGTGCTCTGCCGGCTCTGGTTCAACTGCT GTCCTCTCCGAACG	for hydrophobic core mutants assembly
3AVfor	GGTGCTGTTCCGGCTCTGGTTCAACTGCT GTCCTCTCCGAACG	for hydrophobic core mutants assembly
3AIfor	GGTGCTATCCCGGCTCTGGTTCAACTGCT GTCCTCTCCGAACG	for hydrophobic core mutants assembly
4LLrev	CCACAGAGCTTCTTTCAGCAGCTTCTCGT TCGGAGAGGACAGC	rev hydrophobic core mutants assembly
4LVrev	CCAAACAGCTTCTTTCAGCAGCTTCTCGT TCGGAGAGGACAGC	rev hydrophobic core mutants assembly
4VLrev	CCACAGAGCTTCTTTCAGAACCTTCTCGT TCGGAGAGGACAGC	rev hydrophobic core mutants assembly
4L-rev	CCAAGCAGCTTCTTTCAGCAGCTTCTCGT TCGGAGAGGACAGC	rev hydrophobic core mutants assembly
4-Lrev	CCACAGAGCTTCTTTCAGGATCTTCTCGT TCGGAGAGGACAGC	rev hydrophobic core mutants assembly
5L--for	CTGAAAGAAGCTCTGTGGGCTCTGTCTAA CCTGGCTTCTGGTGGTTGAG	for hydrophobic core mutants assembly
5V--for	CTGAAAGAAGCTGTTTGGGCTCTGTCTAA CCTGGCTTCTGGTGGTTGAG	for hydrophobic core mutants assembly
5-V-for	CTGAAAGAAGCTGCTTGGGTTCTGTCTAA CCTGGCTTCTGGTGGTTGAG	for hydrophobic core mutants assembly
5LV-for	CTGAAAGAAGCTCTGTGGGTTCTGTCTAA CCTGGCTTCTGGTGGTTGAG	for hydrophobic core mutants assembly
5VV-for	CTGAAAGAAGCTGTTTGGGTTCTGTCTAA CCTGGCTTCTGGTGGTTGAG	for hydrophobic core mutants assembly
5--Ifor	CTGAAAGAAGCTGCTTGGGCTCTGTCTAA CATCGCTTCTGGTGGTTGAG	for hydrophobic core mutants assembly
5L-Ifor	CTGAAAGAAGCTCTGTGGGCTCTGTCTAA CATCGCTTCTGGTGGTTGAG	for hydrophobic core mutants assembly
5LVIfor	CTGAAAGAAGCTCTGTGGGTTCTGTCTAA CATCGCTTCTGGTGGTTGAG	for hydrophobic core mutants assembly
6Irev	TTCCTGGTACCCTAAGGTCTCAACCACCA GAAGCGAT	rev hydrophobic core mutants assembly and amplification

Table S2 : oligonucleotides used in assembly of single internal modules

Mutant	oligo #1	oligo #2	oligo #3	oligo #4	oligo #5	oligo #6
mut1	cons1F	cons2R	cons3F	4LLrev	5L--for	cons6R
mut2	cons1F	2A-rev	3A-for	4VLrev	5L--for	cons6R
mut3	cons1F	2A-rev	3A-for	4LLrev	5L--for	cons6R
mut4	cons1F	2A-rev	3A-for	4L-rev	5--Ifor	6Irev
mut5	cons1F	2A-rev	3A-for	4L-rev	5-V-for	cons6R
mut6	cons1F	2A-rev	3A-for	4L-rev	cons5F	cons6R
mut7	cons1F	2A-rev	3A-for	4-Lrev	5L-Ifor	6Irev
mut8	cons1F	2AVrev	3AVfor	4LLrev	5L--for	cons6R
mut9	cons1F	2AIrev	3AIfor	4LLrev	5L--for	cons6R
mut10	cons1F	2A-rev	3A-for	4LVrev	5V--for	cons6R
mut11	cons1F	2A-rev	3A-for	4VLrev	5LV-for	cons6R
mut12	cons1F	2A-rev	3A-for	4LLrev	5LV-for	cons6R
mut13	cons1F	2AIrev	3AIfor	4LLrev	5LV-for	cons6R
mut14	cons1F	2A-rev	3A-for	4LVrev	5VV-for	cons6R
mut15	cons1F	2A-rev	3A-for	cons4R	cons5F	cons6R
mut16	cons1F	cons2R	cons3F	4LLrev	5L-Ifor	6Irev
mut17	cons1F	2AVrev	3AVfor	4LLrev	5LVIfor	6Irev
mut18	cons1F	2AIrev	3AIfor	cons4R	cons5F	cons6R
mut19	cons1F	cons2R	cons3F	4L-rev	cons5F	cons6R

Table S3

Table S3 shows the summary of the results from the simulated annealing approach for the 432 mutants (left column) and the original structures (in the upper part).

In the top rows, energy values for the original crystal structures after simulated annealing are indicated. 2bct, 1q1t, 1ee4 are the original structures of mouse β-catenin, mouse importin-α and yeast importin-α, respectively. Catm, impm, impy are the starting models derived from the original structures where the capping repeats have been replaced by the designed capping repeats Ncap and Ccap.

In the leftmost column (#), each hydrophobic core mutant is identified by a number, from 0 to 431 and the mutants are ordered according to the total rank (second leftmost column). These values do not take into account electrostatic contributions. The value used for the ranking (m1m) is the sum of the ranks of median (me), first percentile (1) and minimum (mi) in all the three starting structures. A second value (m1) is given, which denotes the sum of median and first percentile among the three structures. Individual ranks are also reported for median (me), first percentile (1), minimum (mi) referring to each starting structure. On the right of the ranks, the corresponding potential energy values (expressed as kcal/mol) are indicated for median, first percentile (1^{st} perc.) and minimum in each structure.

For each mutant the hydrophobic core residues are indicated as a change, compared to the original C-type consensus: the dashes indicate no change. The amino acids are indicated in single letter code. The positions not mutated are indicated in gray.

The average volume for core residues of internal repeats is indicated for mouse β-catenin, mouse importin-α, yeast importin-α (corresponding PDB ID are indicated) and for the C-type consensus. In the case of mutants, the core volume is expressed as difference to the core volume of C-type consensus sequence: V(mut)-V(cons). Volumes were calculated according to Chothia [1] in units of $Å^3$.

The mutants selected for expression and characterization are indicated left of the first column with the corresponding name. The first eight mutants, with a ranking value better than the C-type consensus (#264) were selected. Other high ranking mutants, even if not present in the very top group, were selected because of particularly interesting sequences or results in the ranking process.

mut5 (#2): moderate volume increase, comparison with #8, same composition, only position 27 and 28 exchanged.

mut6 (#0): equivalent to consensus among the first 50 top mutants. Always good rank, only high median in importin mouse structure; it was thus considered to be a good candidate.

mut7 (#53): most similar to importin consensus sequence

mut10 (#8): moderate volume increase, comparison with #8, same composition, only position 27 and 28 exchanged. mut13 (#78): high volume and several β-branched residues.

mut14 (#10): high volume and several β-branched residues.

mut15 (#48): simplest mutation with volume gain (G->A) compared to consensus.

mut16 (#221): good rank, only high median in catenin mouse structure; it was thus considered to be a good candidate.

mut17 (#151): high volume and several β-branched residues.

mut18 (#120): good rank, only high median in catenin mouse structure; it was thus considered to be a good candidate.

mut19 (#216): highest ranking mutant with low core volume.

References

1. Chothia, C. (1975). Structural invariants in protein folding. *Nature* **254**, 304-8.

Refinement of consensus designed proteins

The protein YM_4A was characterized as a stable protein with native-like properties, by size exclusion chromatography, ANS binding, circular dichroism and NMR. YM_4A forms a compact structure as indicated also by heteronuclear NOE measurements (Fig. 5): a measured value close to 1 for the transfer of magnetization between a backbone nitrogen and the covalently attached hydrogen indicates a rigid local environment.

The NMR data reveal also an influence of the pH on the HSQC spectrum. Sharp and dispersed peaks are characteristic of well defined and rather rigid structures. The peak broadening and the loss of resolution at pH lower than 10 indicate that protonated groups are involved in the flexibility of the polypeptide (Fig. 6). When the pH is higher than 10, the residues are probably deprotonated and the molecule becomes more rigid.

The stability of YM_4A under different pH conditions was analyzed and is reported in Fig. 7. The CD spectrum is not affected by the shift in pH from 8 to 11. An increase of 4-5°C in the midpoint of denaturation is, however, observed, indicating a higher thermal stability, despite the loss of renaturation efficiency at pH 11.

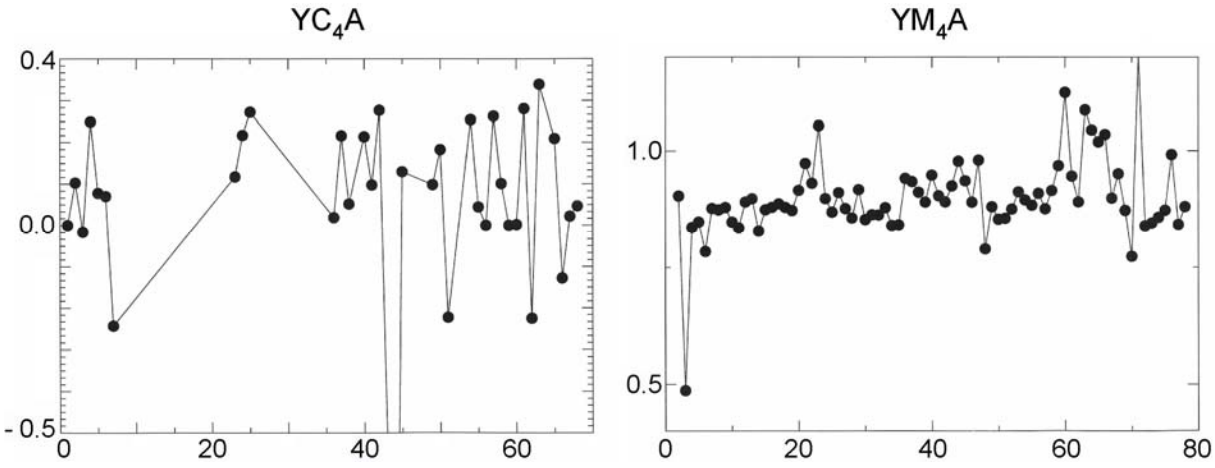


Fig. 5 Heteronuclear NOE signals of YC_4A and YM_4A. The arbitrary values of HetNOEs on the y axis indicate the backbone flexibility of a specific amino acid. Values close to 1 are interpreted as rigid backbone signals, values below 0.6 and negative values are indications of high flexibility. YM_4A appears as a fairly rigid protein, even if the data collection at pH 10 does not allow the detection of all the possible peaks. In contrast, YC_4A (data collected at pH 6) is extremely flexible. The residues are numbered arbitrarily (no assignment was made) and indicated on the x axes. Only the clearly detectable peaks were used for the HetNOE signals determination.

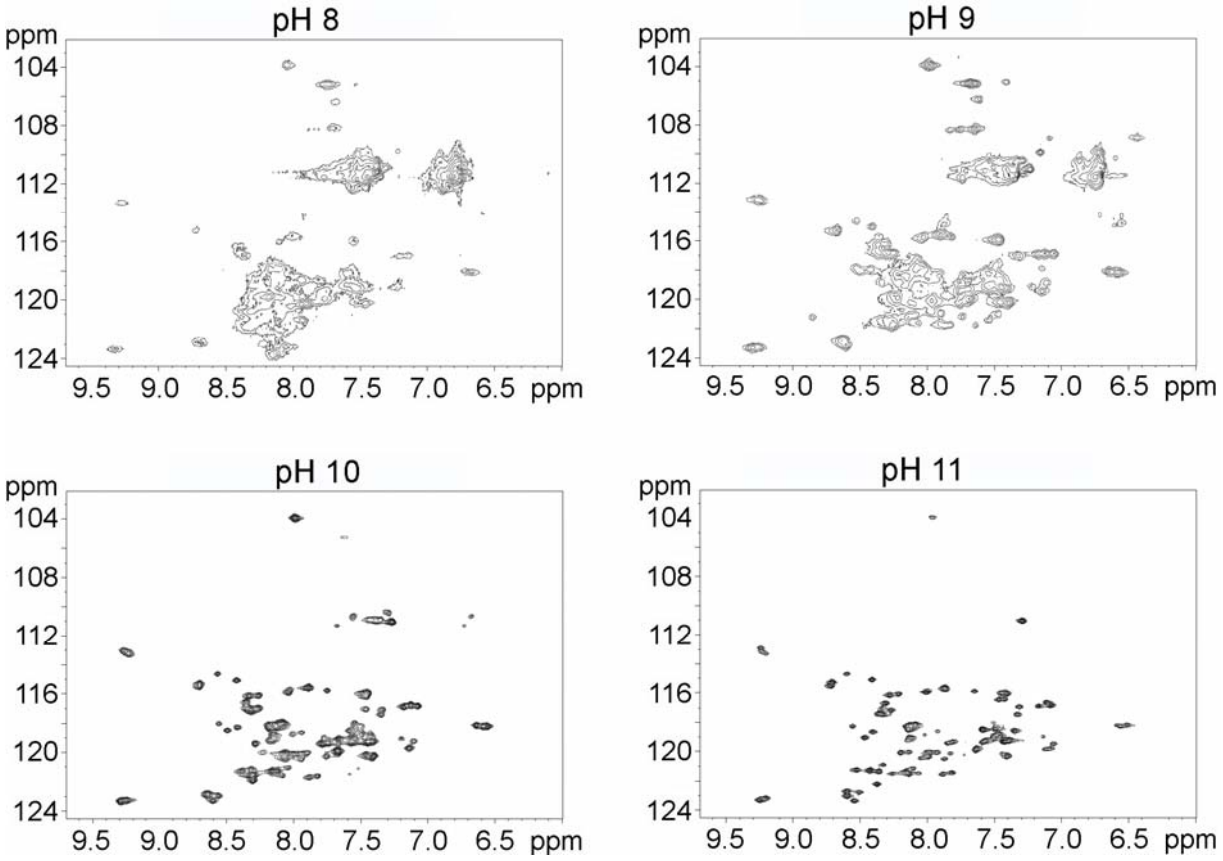


Fig. 6 HSQC spectra of YM_4A at different pH. While increasing the pH value from pH 8 to pH 11, the peaks become more sharp and resolved.

The protonation state is thus really affecting the stability and a possible solution could be the replacement of the residues responsible for this effect. Only Arg and Lys possess a value of protonation constant (pKa=12.5 and 10.8 respectively) compatible with the transition. The two Arg of the molecule are located in the N-terminal capping repeat (another Arg is present as second residue before the His-tag) but their polar groups are probably only marginally involved in interactions with the rest of the molecule, judging from the homology models (Fig. 8a). Two Lys, in contrast, are present in each repeat at positions 26 and 29 and are in close contact in the three-dimensional models. A repulsive effect of the positive charges might account for the destabilization of a compact structure, which disappear when the two Lys, or at least one, become deprotonated. Six more Lys are present in the C-terminal capping repeat but also in this case their interaction with the rest of the structure is less likely than for the internal ones.

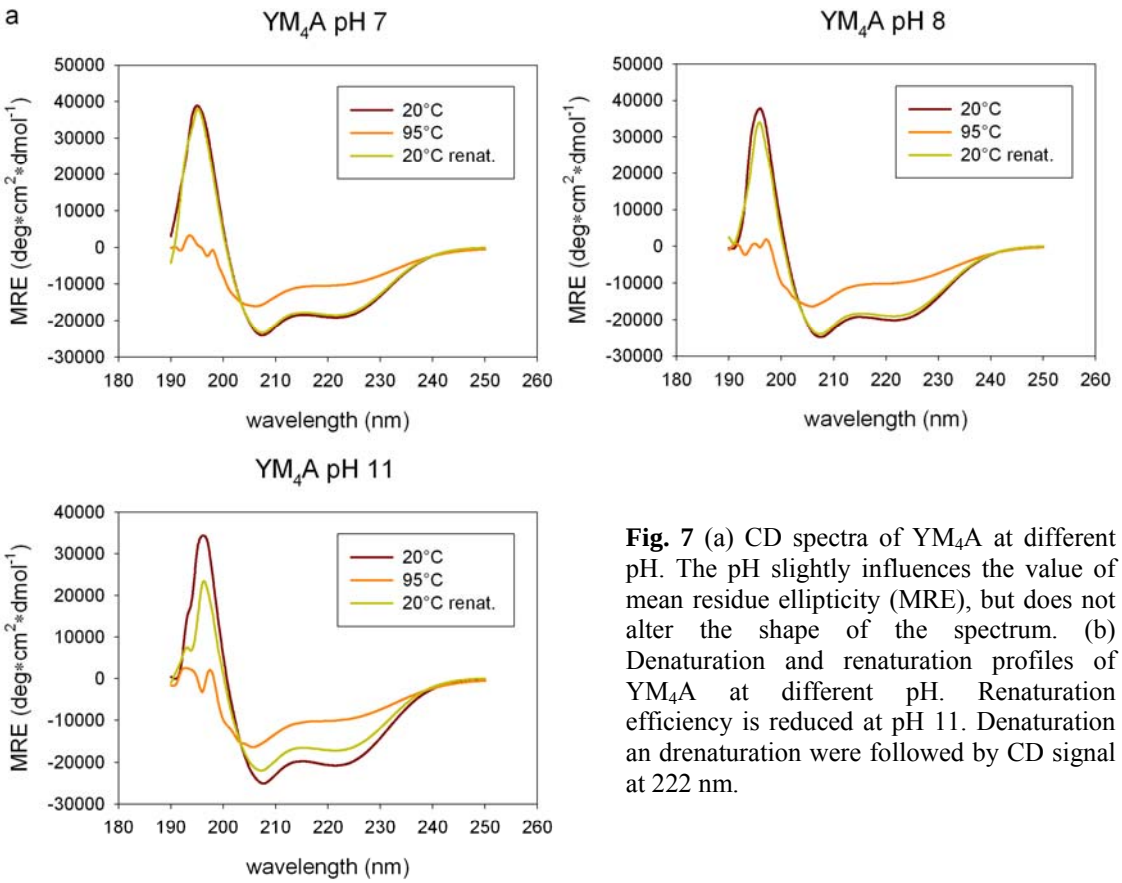
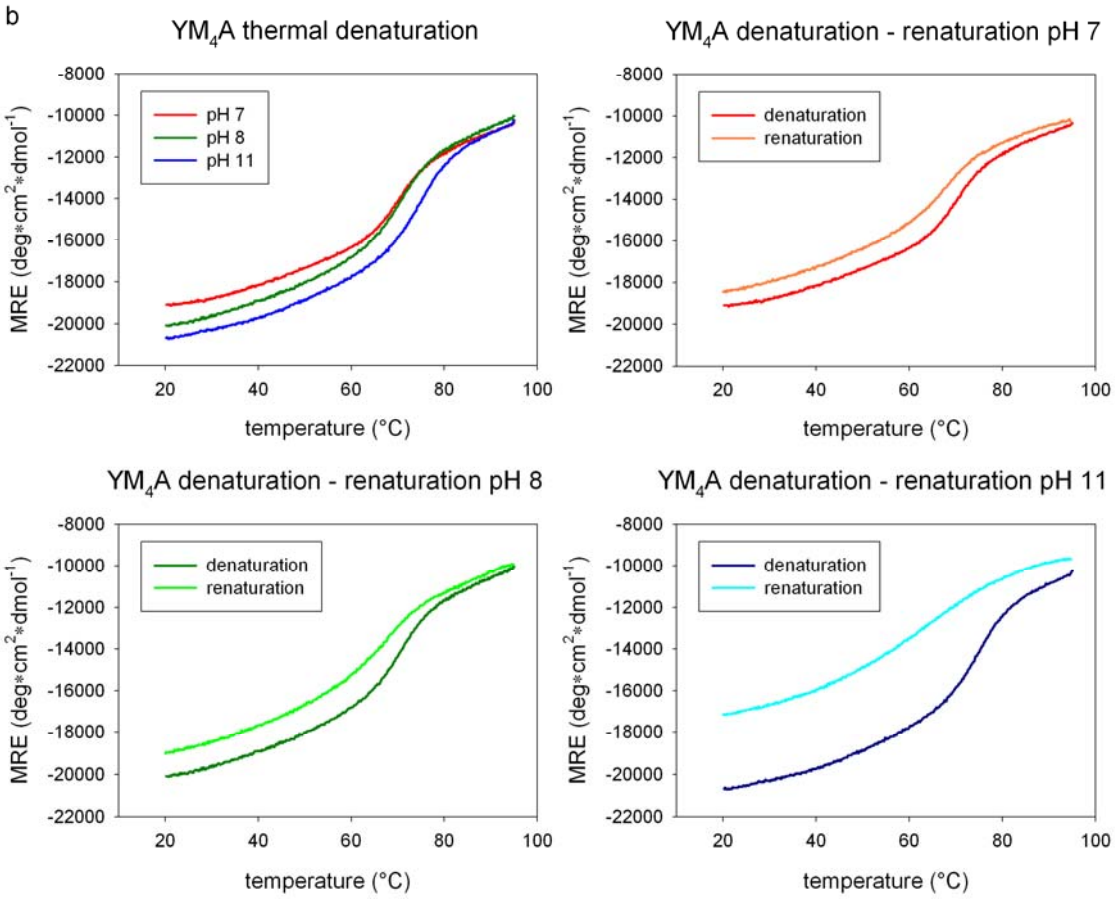
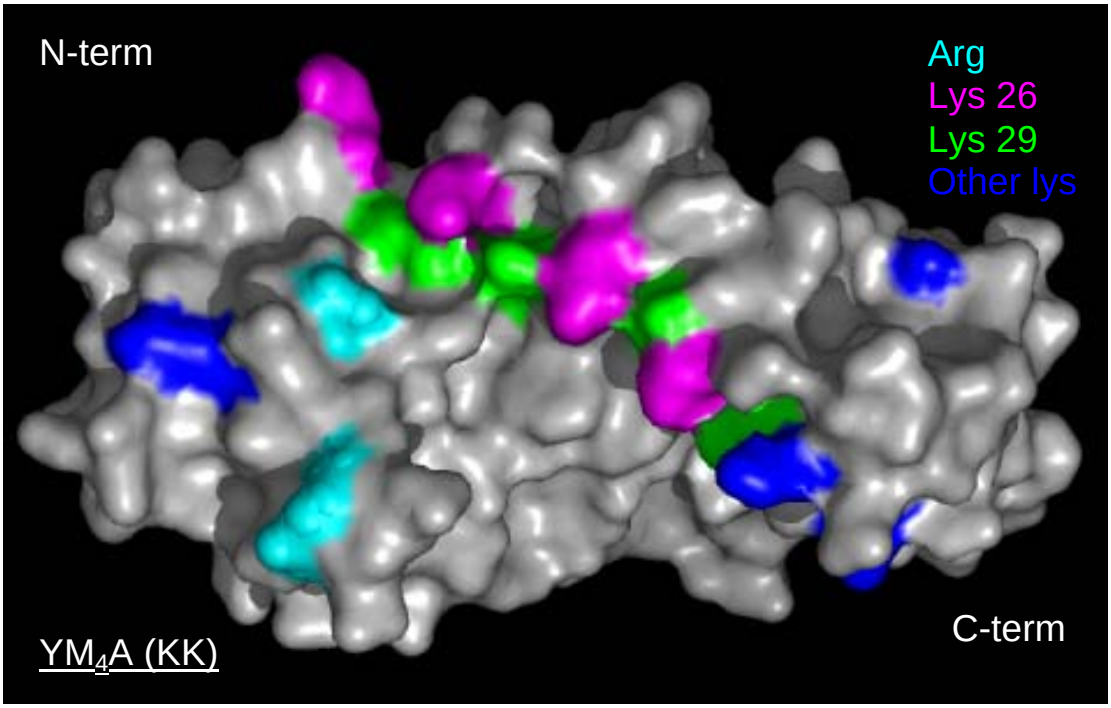


Fig. 7 (a) CD spectra of YM_4A at different pH. The pH slightly influences the value of mean residue ellipticity (MRE), but does not alter the shape of the spectrum. (b) Denaturation and renaturation profiles of YM_4A at different pH. Renaturation efficiency is reduced at pH 11. Denaturation an drenaturation were followed by CD signal at 222 nm.



a



b

```
Ny                  ELPQMTQQLNSDDMQEQLSATVKFRQILSRDG
KK    NEQIQAVIDAGALPALVQLLSSPNEKILKEALWALSNIASGG
KQ    NEQIQAVIDAGALPALVQLLSSPNEKILQEALWALSNIASGG
QK    NEQIQAVIDAGALPALVQLLSSPNEQILKEALWALSNIASGG
QQ    NEQIQAVIDAGALPALVQLLSSPNEQILQEALWALSNIASGG
Ca    NEQKQAVKEAGALEKLEQLQSHENEKIQKEAQEALEKQFSH
```

Fig. 8 Residues involved in pH-dependent stability of YM_4A. (a) Model of YM_4A with residues potentially affecting the change in stability between pH 9 and pH 11. (b) Sequences of capping repeats (Ny and Ca). Arginines and lysines are indicated with the same colors as in Fig. 7a. The mutants called QQ, KQ, QK carry the mutations of Lys->Gln at positions 26 or 29. KK indicates the module M of which YM_4A is built.

To test this hypothesis, three mutant versions of YM_4A were designed with one or two Lys replaced by a Gln. Gln is a polar residue, as required at the protein surface, it possess a long side chain typical of residues found at these positions and can potentially form stabilizing hydrogen bonds with the neighboring residues. In the further description KQ will stand for

mutation K29Q (Lys at position 26 being unaffected), QK for mutation K26Q (Lys at position 29 being unaffected) and the double mutant QQ for K26Q K29Q (Fig. 8b). The original YM_4A will, thus, be referred as KK. Each repeat of the protein will carry the same mutation.

The point mutations were introduced in the single modules by standard mutagenesis protocols (Quickchange, Stratagene) and the modules assembled as described to form proteins with four identical internal modules. The mutants were expressed as soluble proteins and purified with yields up to 50 mg per liter of bacterial colture (Fig. 9). The mass of the proteins was confirmed by mass spectrometry. The characterization, performed as described above by size exclusion chromatography (SEC), multi angle light scattering (MALS), circular dichroism (CD), thermal denaturation and ANS (1-anilino-naphthalene-8-sulfonate) binding is reported in Fig. 10.

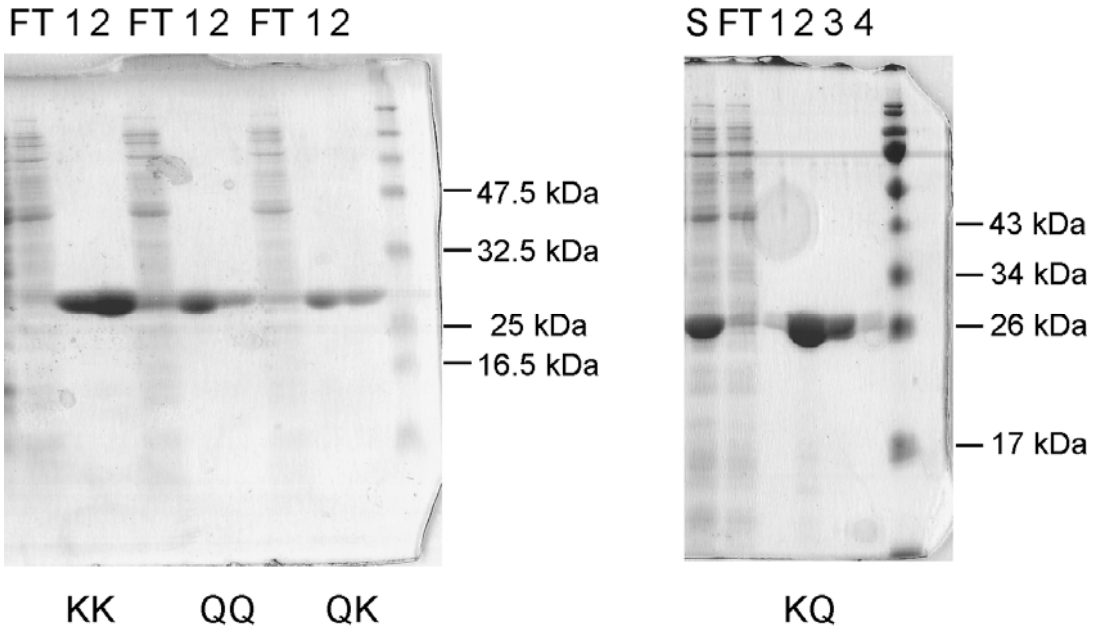


Fig. 9 Expression and purification of YM_4A mutants. S indicates soluble fraction and FT the flow through, 1 to 4 are the eluted fraction from IMAC purification. KK indicates the original YM_4A, QQ, QK, KQ the mutants. The expected size for all the proteins is approximately 27 kDa.

At pH 8 the mutants behave similarly to YM_4A, with the exception of QK that forms oligomers and soluble aggregates eluting at high molecular weight in SEC (Fig. 10a). The correct size of the other proteins was confirmed by MALS (Fig. 10a). The values observed in CD (Fig. 10b) and ANS binding (Fig. 10c) are, however, very similar for all the proteins. The aggregated QK does not bind ANS significantly, indicating that the hydrophobic core is protected from the solvent access. All the mutants show an increase in thermal stability of

approximately 8°C compared to YM_4A, when considering the midpoint of transition for qualitative comparison (Fig. 10d). Therefore, the mutations introduced stabilize the protein, but one combination (QK) is aggregation-prone. Compared to the other mutants, QK possess a higher value in mean residue ellipticity and a stable pre-transition baseline in thermal denaturation, both probably due to the aggregated form, which breaks up around 50°C. Above 50°C the profile is comparable to the other proteins.

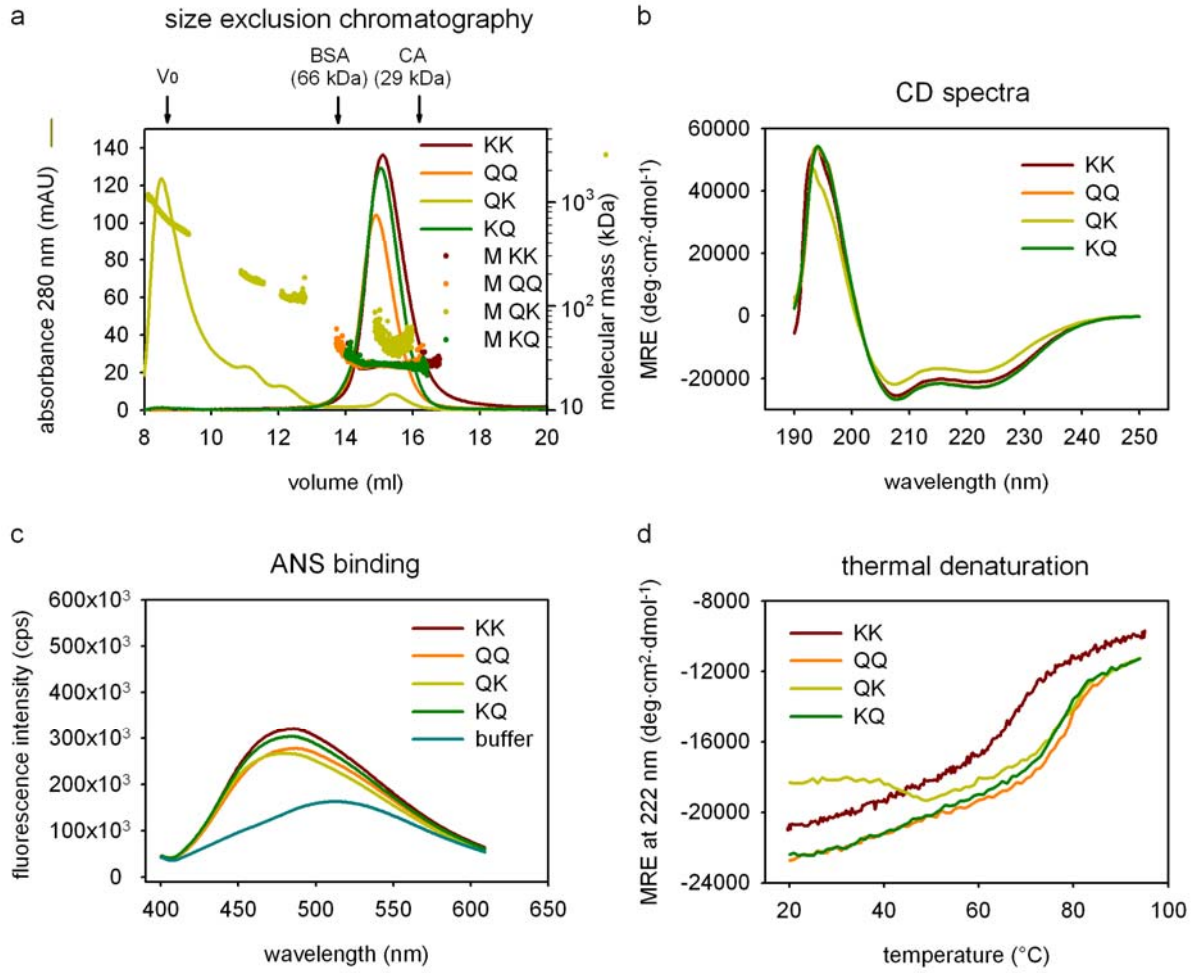


Fig. 10 Biophysical characterization of KK and mutants. (a) size exclusion chromatography with Superdex 200 column and light scattering. QK shows aggregation and two types of oligomers in addition to a small monomeric fraction. The other mutants behave as KK. The molecular mass, indicated by M, was calculated using MALS. V0 indicates the void volume, BSA (bovin serum albumin) and CA (carbonic anhydrase) are the molecular weight standards. (b) CD spectra of the mutants are very similar to KK: QK shows a small loss in ellipticity and the other two versions an increase. (c) Thermal denaturation followed by CD at 222 nm. Unfolding curves possess the same profile but mutants have a midpoint of denaturation increased by approximately 8°C. (d) ANS binding signals are similar among all KK and the mutants. All the measurements were performed at pH 8.

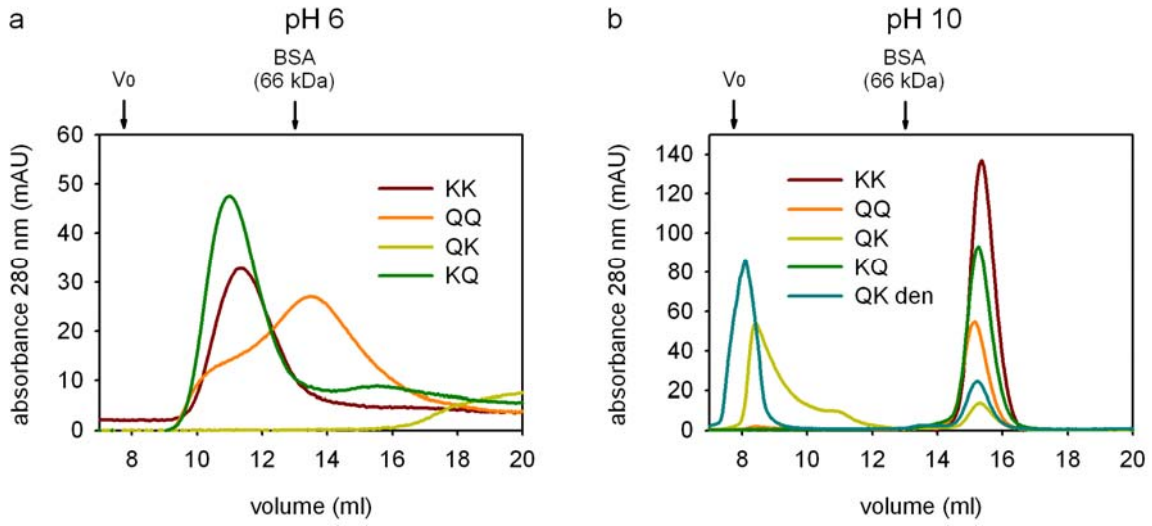


Fig. 11 pH dependent SEC of mutants. At pH 6 (a) aggregation and early elution are observed for YM_4A and the mutants. QK interacts with the column and elutes only gradually. At pH 10 (b) the elution profile are similar to the what observed at pH 8, with a monomeric peak for KK, KQ, QQ and a main aggregated fraction for QK. SEC was performed with 20 mM MES (2-N morpholino ethanesulfonic acid), 50 mM NaCl (pH 6) and 20 mM CAPS ((3-cyclohexylamino)-2-hydroxy-1-propanesulfonic acid), 50 mM NaCl (pH 10).

Additional SEC experiments were performed at pH 6 and pH 10 to verify the tendency of aggregation of the mutants (Fig. 11). The mutants behave similarly to YM_4A, forming aggregates at low pH and remaining monomeric at high pH. Surprisingly, at pH 6 QK does not even elute but interacts with the column. A process of thermal denaturation and renaturation could potentially break down the aggregates and increase the fraction of monomeric QK, already present in low amounts. The protein was tested by SEC at pH 10 after renaturation at pH 8, but, in contrast to the original hypothesis, the soluble aggregates were found again as the dominating species (Fig. 11b).

The proteins KQ, QK and QQ were further analyzed by HSQC NMR experiments at a concentration of 0.5 mM. The major peak of soluble aggregates was used in the case of QK. The data were collected at pH 8, pH 9 and pH 11 as described previously.

When comparing the spectra at pH 9 with the spectrum of YM_4A (Fig. 12), the resolution of the peaks appears clearly improved, similar to the spectrum of YM_4A at pH 11 but with more visible peaks due to several signals that disappear at high pH. The mutants seem to be structurally more rigid at pH 9 than YM_4A, as inferred already from the denaturation curves.

Additional HetNOE experiments confirmed the HSQC results. The NOE signals, close to the value of 1 at pH 9, indicate that the detected residues of both KQ and QQ do not experience backbone flexibility (Fig. 13), similarly to what observed for YM_4A (Fig. 5). The signal of QK decreases over time, probably as results of ongoing aggregation processes, and

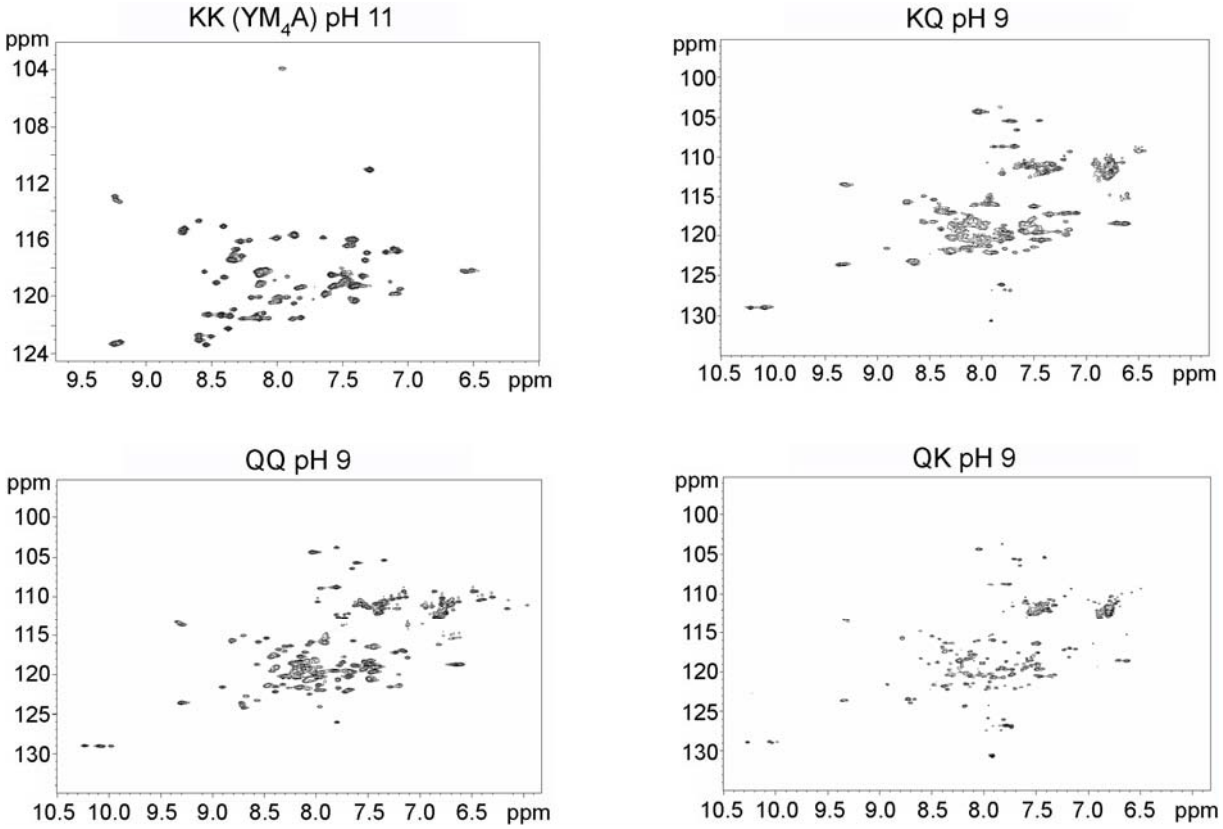


Fig. 12 Comparison of HSQC spectra. The spectra observed at pH 9 for the mutants are similar to the result obtained with YM_4A at pH 11.

the sample could not be used for HetNOE measurements. The results support the initial hypothesis of stability being influenced by Lys charge repulsion. However, a number of residues at pH 8 in KQ and QQ show a significant decrease of NOE values in comparison to pH9, suggesting loss of rigidity for several residues. Unfortunately, assignment of the signals to the protein residues is problematic, due to the size limitations (MW 27 kDa) and the repeated sequence, preventing the identification of the flexible residues and their effects on the structure.

However, the reduction of the pH-dependence of the protein stability from pH 11 to pH 9 represents a significant improvement that brought the working range of the scaffold to more standard conditions for proteins. The armadillo proteins containing modules with the double lysine (M module) are still extremely valuable, for their ability of resisting to high pH conditions and for the advantages they could provide in terms of electrostatic interaction in case of negatively charged peptides. Therefore, the next step will be the generation of libraries based on these protein scaffolds.

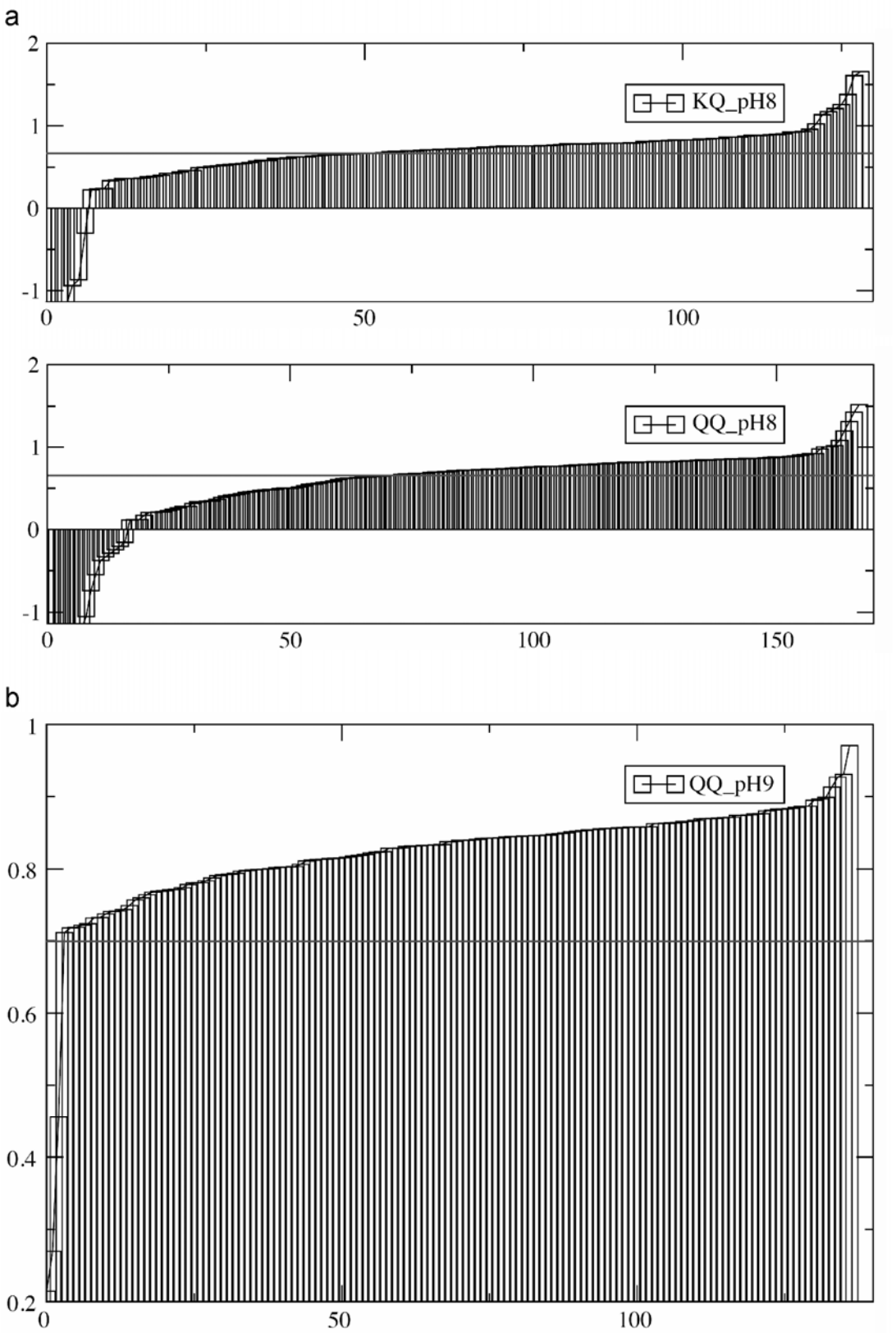


Fig. 13 HetNOE measurements of mutants. The HetNOE signals of up to 150 residues could be measured, in contrast to a maximum of 80 for YM_4A at pH 11. The orizontal line represents an indicative value of 0.7 to qualitatively discriminate between residues with rigid or flexible backbone. (a) At pH 8 more than 60% of the signals belong to residues with rigid backbone, for both KQ and QQ. Lower values indicate a flexible backbone. (b) At pH 9 almost all the signals in QQ are above 0.7. QK was not measured due to precipitation and reduction of the concentration in solution during the experiment.

References

1. Nusslein-Volhard, C. & Wieschaus, E. (1980). Mutations affecting segment number and polarity in Drosophila. *Nature* **287**, 795-801.
2. Wieschaus, E., Nüsslein-Volhard, C. & Jürgens, G. (1984). Mutations affecting the pattern of the larval cuticle in drosophila melanogaster .3. Zygotic loci on the X-chromosome and 4th chromosome. *Wilhelm Rouxs Archives of Developmental Biology* **193**, 296-307.
3. Riggleman, B., Wieschaus, E. & Schedl, P. (1989). Molecular analysis of the armadillo locus: uniformly distributed transcripts and a protein with novel internal repeats are associated with a Drosophila segment polarity gene. *Genes Dev* **3**, 96-113.
4. Peifer, M., Berg, S. & Reynolds, A. B. (1994). A repeating amino acid motif shared by proteins with diverse cellular roles. *Cell* **76**, 789-91.
5. Huber, A. H., Nelson, W. J. & Weis, W. I. (1997). Three-dimensional structure of the armadillo repeat region of beta-catenin. *Cell* **90**, 871-82.
6. Conti, E., Uy, M., Leighton, L., Blobel, G. & Kuriyan, J. (1998). Crystallographic analysis of the recognition of a nuclear localization signal by the nuclear import factor karyopherin alpha. *Cell* **94**, 193-204.
7. Andrade, M. A., Petosa, C., O'Donoghue, S. I., Müller, C. W. & Bork, P. (2001). Comparison of ARM and HEAT protein repeats. *J Mol Biol* **309**, 1-18.
8. Finn, R. D., Mistry, J., Schuster-Bockler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., Moxon, S., Marshall, M., Khanna, A., Durbin, R., Eddy, S. R., Sonnhammer, E. L. & Bateman, A. (2006). Pfam: clans, web tools and services. *Nucleic Acids Res* **34**, D247-51.
9. Sambrook, J. & Russell, D. W. (2001). *Molecular cloning : a laboratory manual*. 3rd edit, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.
10. Schultz, J., Milpetz, F., Bork, P. & Ponting, C. P. (1998). SMART, a simple modular architecture research tool: identification of signaling domains. *Proc Natl Acad Sci U S A* **95**, 5857-64.
11. Malik, H. S., Eickbush, T. H. & Goldfarb, D. S. (1997). Evolutionary specialization of the nuclear targeting apparatus. *Proc Natl Acad Sci U S A* **94**, 13738-42.
12. Andrade, M. A., Perez-Iratxeta, C. & Ponting, C. P. (2001). Protein repeats: structures, functions, and evolution. *J Struct Biol* **134**, 117-31.
13. Coates, J. C. (2003). Armadillo repeat proteins: beyond the animal kingdom. *Trends Cell Biol* **13**, 463-71.
14. Edwards, T. A., Pyle, S. E., Wharton, R. P. & Aggarwal, A. K. (2001). Structure of Pumilio reveals similarity between RNA and peptide binding motifs. *Cell* **105**, 281-9.
15. Milburn, C. C., Boudeau, J., Deak, M., Alessi, D. R. & van Aalten, D. M. (2004). Crystal structure of MO25 alpha in complex with the C terminus of the pseudo kinase STE20-related adaptor. *Nat Struct Mol Biol* **11**, 193-200.
16. Rose, R., Weyand, M., Lammers, M., Ishizaki, T., Ahmadian, M. R. & Wittinghofer, A. (2005). Structural and mechanistic insights into the interaction between Rho and mammalian Dia. *Nature* **435**, 513-8.

Chapter 3
Libraries of Designed Armadillo Repeat Proteins

A library for peptide binding

Many scaffolds for protein recognition have been generated in the last decades as alternative to antibodies [1; 2; 3]. One of the aims was to obtain libraries with the broadest possible range, able to provide binders against any type of target. Peptides have not been a preferential target and very few libraries were designed to provide specific binders, notably based on PDZ [4] and SH2 [5] domains. The main goal was to investigate the specificity of these domains and to provide molecules able to interfere with cellular recognition processes. Theses domains are quite restricted however in the range of targets they are able to recognize: C-terminal peptides for PDZ and phosphor-tyrosine peptides for SH2. No library for the recognition of peptides of arbitrary sequence has been then designed so far.

A scaffold based on armadillo repeat proteins is not restricted to particular target sequences. The recognition of the backbone grants a constant binding mode, strengthened by the interaction with the side chains taking place at the helix 3 surface. No post-translational modification or free terminus is required. The only small limitation is represented by proline. The backbone nitrogen, involved in the formation of the pentameric ring, cannot be involved in any hydrogen bond, weakening the backbone binding.

The modular structure of designed armadillo repeat proteins can provide a dipeptide-specific recognition, taking advantage of the regularity of the binding site. Even if potentially accessible to all the designed repeat proteins, such modularity cannot be fully exploited if the targets, or part of them, cannot be recognized in a conserved and general way.

Having the scaffold, the next step is the generation of the library that will be used for the selection of the binders. Several aspects come together when designing a library, first of all the structural constraints depending on the scaffold, but also the general concepts encompassing the library size, its diversity and the methods to achieve the desired level of randomization.

The concept of library

The properties of libraries, and specifically of proteins libraries, can be rigorously treated using mathematical expressions. An in depth view of library description in mathematical terms is provided by Bosley and Ostermeier [6] and the screening process is discussed by Denault and

Pelletier [7]. I will refer here to their work to summarize some of the key concepts in the creation of a library based on armadillo repeat proteins.

Library size (or complexity) refers to the number of different members present. The theoretical complexity is related to the method used to generate the library and indicates the maximum number of different variants. The practical complexity is the number of clones that can be reasonably handled during a selection or screening. The relation between the size of the library used as input for the selection and the capacity of the selection system determines the coverage of the selection, defined as the percentage of the initial library that can be, statistically, analyzed during the selection. As example, a selection method that can handle a number of molecules equal to 20 times the library size has more than 99% of coverage [7], which is the probability that every variant has been picked at least once. High coverage values are possible only for small libraries. In case of new binding proteins, large libraries are often required to select molecules with new specificities. A small library with a restricted set of mutations could completely avoid whole families of potential binders, which will be present in a large library, even thought the best binders could be missed. A strategy combining a low coverage and further affinity maturation of the selected molecules would represent in this case the preferred choice.

A library size, however, depends on the methods used to generate the complexity [8]. For error prone PCR the theoretical complexity can be estimated knowing the polymerase error rate and the length of the gene. Tailor-made libraries, generated by using oligonucleotides with randomized positions, allow the calculation of the complexity, provided that the ratio between the different nucleotides is known. An even more controlled variability can be generated using trinucleotide phosphoramidites [9]. The insertion of pre-synthesized trinucleotides allows the introduction of the desired mixture of amino acids without the risk of inserting stop codons or undesired amino acids. The ratio between amino acids can be also controlled by the amount of specific trinucleotides employed during the oligonucleotide synthesis, leading to a better quality of the library.

In general, an indication of library quality is the number of functional members. Frameshifts and stop codons, as results of mutagenesis, PCR or other manipulations, lead to truncated non-functional proteins. Ligation into a vector is never 100% efficient and, when a transformation step is involved, a fraction of the resulting clones will not contain any gene. Transformation is also a major source of library size reduction, due to its limited capacity. Methods that do not require any transformation step have a capacity of two to three orders of

magnitude higher than methods that require it (e.g. phage display and yeast display). If the introduction of stop codons can be calculated in case of oligonucleotides with randomized positions, the transformation efficiency and the fraction of empty vectors are usually determined experimentally. Statistics based on sequencing results finally provides an overview on the overall quality of a library, in terms of desired and unwanted mutations, expected amino acid ratios and fraction of functional proteins. The frameshifts observed are usually due to mistakes in the oligonucleotides.

An additional “curing” of the library can be provided by in-frame selection systems, where proteins necessary for the survival of cells are expressed as C-terminal fusion proteins of the library members. Only members that are correctly translated lead to the production of the second protein and to the survival of the cells. A system recently developed uses a split intein as alternative for the selection [10]. Depending on the library, the advantage of an in-frame selection must be, however, carefully balanced with the requirement of an additional ligation in a suitable vector and a transformation step that will probably reduce the library size.

The armadillo library

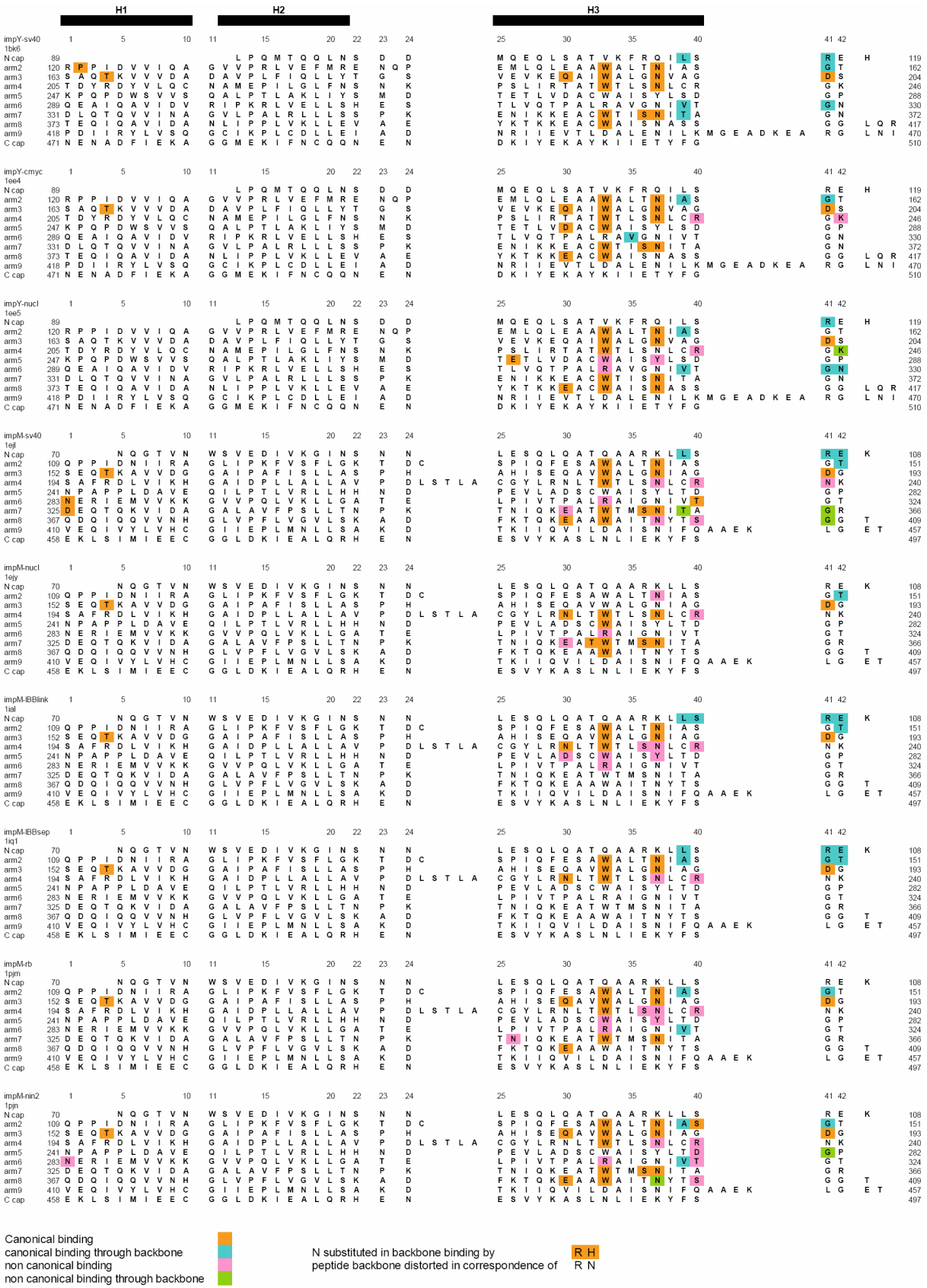
Positions responsible for binding can be identified from the crystal structures of complexes of armadillo repeat proteins with their targets. The results of the analysis are reported in Fig. 1. All the positions involved, except position 41, are located on helices and among them one on helix H1 and the others on helix H3. Recognition by secondary structure elements is characteristic of solenoid-like repeat proteins, in contrast to antibodies and other alternative scaffolds where the binding site is mainly formed by loops. The rigidity of these elements allows, in the armadillo proteins, the formation of the desired general binding mode that we would like to obtain in our library.

Asn in position 37 is responsible for the recognition of the target backbone and the general binding mode; it is already present in the consensus and it will be kept constant. Position 4 contributes both to the target binding and to the hydrophobic core formation; the residues at this position will be limited to the few compatible with the core packing. The residues appearing in natural proteins have been included, with the exception of Pro, which could destabilize the helix H1 and Ala, which is too small to reach the side chain of the targets. Glu has been added for the potential ability to interact with positive charges, even if not present in

the list of natural residues. Positions 26, 29 and 30 are used alternatively for the recognition of long target side chains: position 29 in catenin subfamily, position 30 and once position 26 in importin subfamily. Among them, only position 30 has been selected for randomization, being the most frequently used and considering that the importin binding is generally more regular than the catenin binding mode. Position 33 contains a Trp in the consensus and in the importin subfamily. The space occupied allows almost any residue at this position. Position 36 shows high variability, even though small residues are preferred in importins. Residues at this position could contribute significantly to binding. However, several residues could disrupt the backbone binding of the neighboring Asn37, replacing it, as observed in some catenin complexes, in the role of target main chain binder. Position 40 shows high variability, but large side chains are often present. Position 41 is part of the loop connecting H3 with H1 of the contiguous repeat. It interacts with the side chains of the target peptide mainly by backbone hydrogen bonds. However, several residues could be accommodated at this position providing new types of interaction. An almost complete randomization (no Cys to avoid disulfide bond formation, no Gly or Pro that could possibly disrupt the helix H3 without providing binding advantages) was the strategy adopted for all the positions except position 4, where the residues were limited to Glu, His, Lys, Arg, Ile, Gln, Thr.

Possible effects of the randomization on the helices were tested *in silico* using the program AGADIR [11; 12], originally developed to predict helical propensity of peptides in solution. The helices of our modules are not free in solution but the randomized positions are exposed to the solvent and are not expected to interact with the core of the protein. An analysis of the calculated helical propensities for the binding helix H3 will be valuable when compared to the original helix H3 of the module M. A decrease in the helical propensity for several library members will be an indication of an average destabilization of the helix and probably of the whole protein. AGADIR is extremely sensitive to helix length and to capping residues. Calculations were done in presence or absence of the first and last residue not expected to be part of the α-helix and compared. The absolute values of helical propensity are different but they show the same distribution in relation to the non-mutated helix.

Fig. 1 (next page) Binding residues in natural armadillo repeat proteins. The protein sequences of importins α and β-catenins are depicted with the repeats aligned according to the structural data. The residues involved in binding, as defined by analysis of crystal structures of complexes, are colored. Orange indicates a canonical binding mode, with the target backbone bound every two residues by an asparagine at position 37 in the armadillo repeat. Cyan indicates that the main chain of that armadillo residue is used for the recognition of a target side chain. Magenta is used for residues contacting the target main or side chain in alternative ways (e.g. pos 36 contacts the backbone instead of position 37), and green when the interaction involves the backbone of a residue of the armadillo domain.


H1
H2
H3
impY-sv40
1bk6
impY-cmyc
1ee4
impY-nucl
1ee5
impM-sv40
1ejl
impM-nucl
1ejy
impM-IBBlink
1ial
impM-IBBsep
1iq1
impM-rb
1pjm
impM-nin2
1pjn
N cap
arm2
arm3
arm4
arm5
arm6
arm7
arm8
arm9
C cap
Canonical binding
canonical binding through backbone
non canonical binding
non canonical binding through backbone
N substituted in backbone binding by
peptide backbone distorted in correspondence of
R H
R N


H1
H2
H3
bcatM-Pcadherin 1i7w
bcatM-cadherin 1i7x
bcatM-APC 1jpp
bcatM-ICAT 1m1e
bcatH-XTcf3 1g3j
bcatH-XTcf4 1jdh
bcatH-XTcf4bis 1jpw
bcatH-ICAT 1luj
bcatH-APC20P 1th1
N cap
arm2
arm3
arm4
arm5
arm6
arm7
arm8
arm9
arm10
arm11
C cap

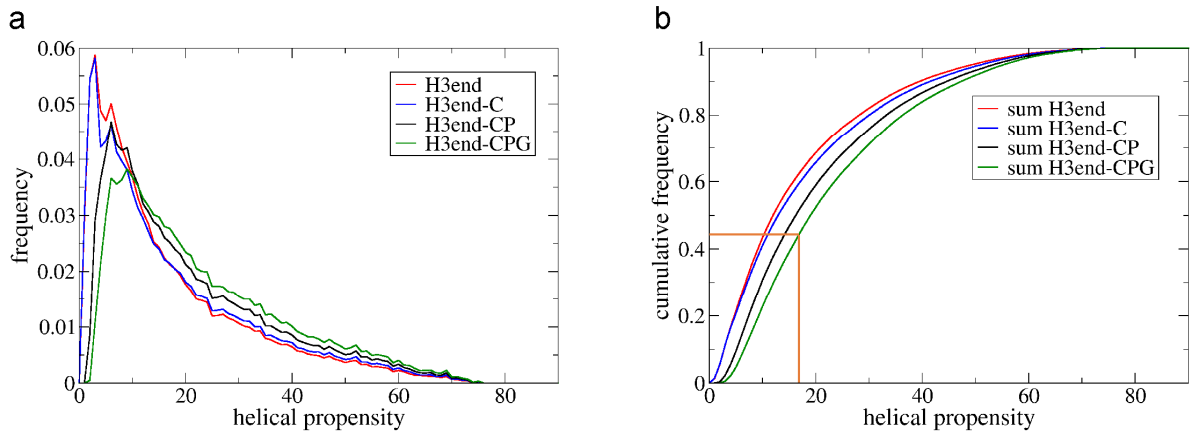


Fig. 2 Helical propensity of KK library members. (a) Frequency of helical propensity value of library members. The calculation of helical propensity was done with the program AGADIR using the helix 3 (positions 25-40) and including one additional residue at N and C termini (Asp24 and Gly41). The colors refer to complete or partial randomization o fthe consensus sequence. The individual members of H3end (red line) were calculated allowing any residue at the randomized positions. For H3end-C Cys was not allowed, similarly for H3end-CP (Cys and Pro excluded) and H3end-CPG (Cys, Gly and Pro excluded). b) Cumulative frequency of helical propensity of library members. The orange line indicates the value corresponding to the non randomized H3. More than 50% of the library members have higher value of helical propensity than the consensus sequence. Helical propensity is indicated as percentage.

Fig. 2 illustrates the results using H3 sequence and capping residues, from position 24 to 41. Going from a complete randomization to a restricted set without Cys, Pro, Gly the frequency of sequences with low helical propensity decreases (Fig. 2a). When comparing the cumulative frequency (Fig. 2b) with the value of unmodified H3 (17.9), more than 50% of the library possesses a higher helical propensity. The situation is similar for the module variants QQ, QK and KQ.

The isoelectric point (pI) distribution of a putative N3C library (N-cap, three internal modules, C-cap) was calculated using the program polygen, written by Andreas Ernst [13]. The library members have pIs shifted toward higher values, compared to the unmodified YM_3A, especially when an N-terminal histidine tag is present, like in the analyzed armadillo repeat proteins (Fig. 3). This behavior is probably due to the presence of more positively (Lys, Arg, His) than negatively charged (Asp, Glu) amino acids among the allowed residues for randomization, and to the loss of one negative charge occupying a randomized position (Glu at position 30) in the consensus sequence.

A library module was assembled from oligonucleotides as previously described, based on the M module. However, due to randomization, the sequence at position 41 could not be used as site for restriction and ligation. The DNA sequence was then shifted by 6 bases compared to the original module M. The capping repeats were shifted accordingly in order to produce, after

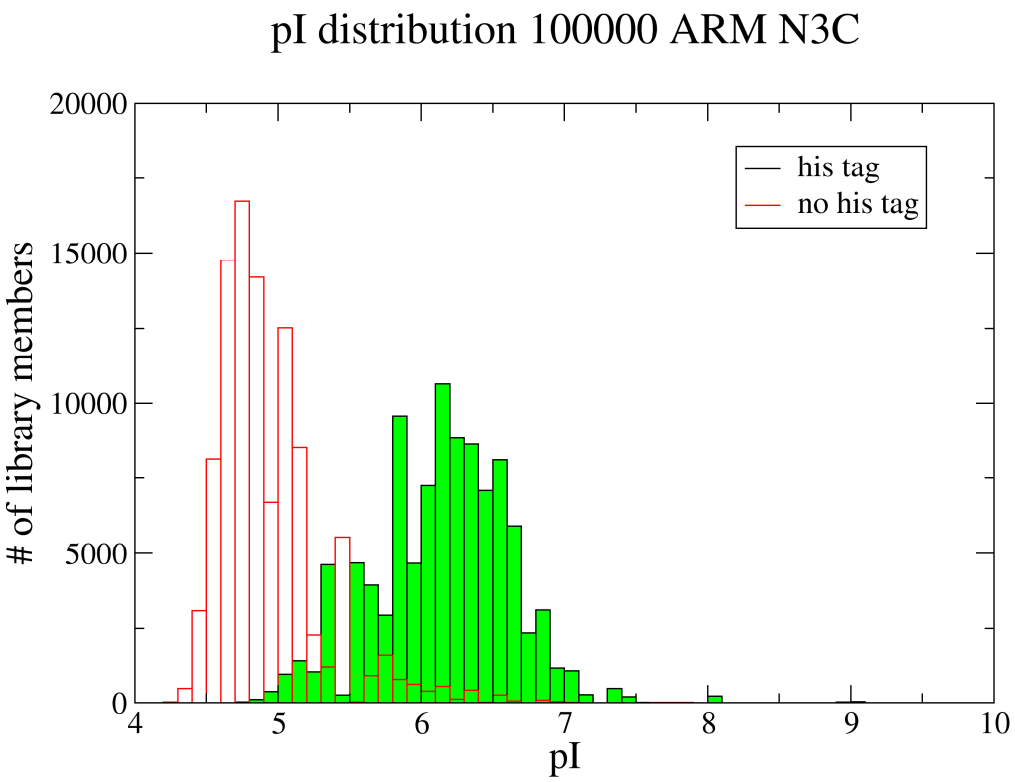


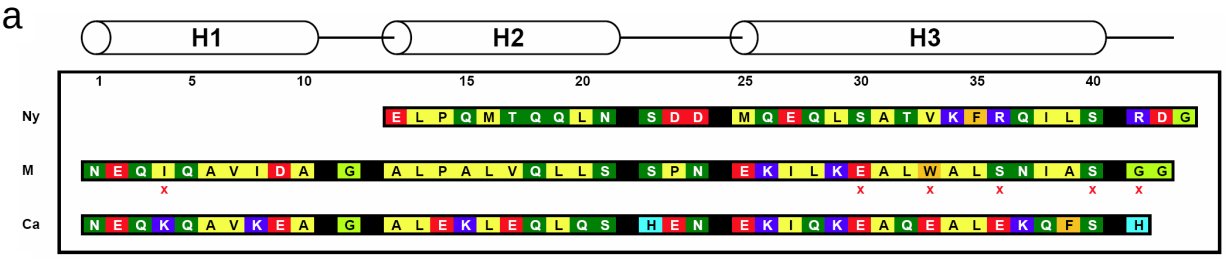
Fig. 3 pI distribution of library members. The pI distribution of the library members is reported in red with no filling. In black with green filling are shown the results when the histidine tag, used for purification, is taken into account. The calculation is based on 100000 sequences randomly selected from the N3C library. The pI of YM_3A, considered as reference is 4.64, without histidine tag and 5.23 with it. The results were obtained using the program polygen, written by Andreas Ernst.

ligation, the correct sequence of designed armadillo repeat proteins. The oligonucleotides used for the assembly of the internal module and the capping repeats are listed in Appendix 3. Position 4 was randomized using a combination of three degenerated oligonucleotides. Only the allowed residues were encoded by the codons and were equally represented. The other positions were randomized using trinucleotide phosphoramidites (Glen Research, USA) inserted in a single oligonucleotide (Metabion, Germany), reaching a theoretical diversity of 9.9×10^6 per module.

The modules were assembled together as described, leading to a final library containing three internal randomized modules. The theoretical diversity of a N3C library would be 9.7×10^{20}, but this value was reduced during the assembly. After each ligation step the concentration of purified ligation product was measured. A sample containing approximately 10^{11} molecule was used as template for PCR. At the first ligation step the maximum number of variants would be 9.8×10^{13}. It is reasonable to assume that 10^{11} molecules would be probably all different and the chance to preferentially ligate some of the sequences is negligible, having

all of them the same length and being treated in the same way. Out of all the possible combination of codons introduced ($9.9x10^6$), only one (Glu-Asp GAA-GAC) at positions 40-41 leads to a recognition site for BpiI that can be cleaved during assembly. Therefore, the library cannot be biased because of the randomized sequences. After each ligation step the practical library size of the N3C library was approximately 10^{11}. This library size is easily handled and *in vitro* selection methods can grant full coverage.

a



b

Position 4: E H K I Q T R

residues	counted (79)	% expected	% observed
E	9	14.3	11.4
H	15	14.3	19.0
K	13	14.3	16.5
I	11	14.3	13.9
Q	9	14.3	11.4
T	11	14.3	13.9
R	11	14.3	13.9

Other randomized positions
(all amino acids except Cys, Gly, Pro)

residues	counted (399)	% expected	% observed
A	27	5.9	6.8
D	17	5.9	4.3
E	27	5.9	6.8
F	21	5.9	5.3
H	34	5.9	8.5
I	31	5.9	7.8
K	21	5.9	5.3
L	14	5.9	3.5
M	24	5.9	6.0
N	15	5.9	3.8
Q	18	5.9	4.5
R	16	5.9	4.0
S	27	5.9	6.8
T	26	5.9	6.5
V	34	5.9	8.5
W	23	5.9	5.8
Y	24	5.9	6.0

Single module
Clone sequenced: 23 repeats

Framework point mutation: 1
Framework point del. and ins. (frameshift): 2
Framework del. and ins. (no frameshift): -
Rand. Pos. point mut.: -
Rand. Pos. point del. and ins. (frameshift): 4
Rand. Pos. del. and ins. (no frameshift): 2
Out of frame sequences: 6
In frame sequences: 17

N3C
Clone sequenced: 20 N3C, 60 repeats

Framework point mutation: 13 (4 silent)
Framework point del. and ins. (frameshift): 7
Framework del. and ins. (no frameshift): 3
Rand. Pos. point mut.: 2 (1 silent, 1 stop)
Rand. Pos. point del. and ins. (frameshift): 2
Rand. Pos. del. and ins. (no frameshift): 3
Out of frame sequences: 7
In frame sequences: 13

Rand. Pos., randomized position
Point mut., point mutation
del., deletion
ins., insertion

Fig. 4 Library quality. The sequences of the capping repeats and the internal module M are indicated in (a), with the randomized position labeled by the red x. (b) Expected and observed frequency of residues at randomized positions, indicated and mutations observed in single modules and members of the N3C library. Only a restricted number of residues was allowed at position 4 because of steric constraints. The number of in frame and out of frame sequences is indicated. Insertions or deletions without frameshift are the result of the insertion or deletion of 3 nucleotides, or more often the codon of one amino acid.

The quality of the library was evaluated by sequencing of randomly picked modules or N3C clones. The frequency of each residue at randomized positions corresponds to the expected values (Fig. 4) and a considerable number of clones were in frame: 17 out of 23 (74%) single modules and 13 out of 20 (65%) N3C armadillo proteins. Following this estimation, the practical diversity of the N3C library can be considered as $6.5x10^{10}$.

Similar libraries, both for single modules and N3C proteins, were generated based on KQ, QQ and QK variants, described in Chapter 2. The Lys->Gln mutations were introduced during module assembly using primers carrying the mutations and amplifying the product of the annealing of oligonucleotides lib5F and lib6R. The primer sequences are listed in Appendix 3. The four libraries will represent, therefore, the source for further *in vivo* or *in vitro* selection.

Characterization of unselected library members

Ten full length unselected library members (1_3, 1_4, 2_1, 2_4, 2_7, 2_8, 2_9, 2_10, 2_12, 2_14) were expressed and characterized analyzing their behavior in SEC, CD and ANS binding, in comparison to the consensus protein YM_3A. 1_3, 1_4, 2_4 do not possess any tryptophan. Their concentration was determined using the method described in Appendix 1, based on absorbance at 235nm and 280 nm. The results are summarized in Table 1.

All the proteins were soluble and expressed with yield (Fig. 5) similar YM_3A. However, all of them, except 1_3 and 2_8, showed reduced binding to the Ni-NTA column material, and were present in the flow through. The purification of YM_3A, in contrast, did not show any residual protein in the flow through. Consistent with this finding, SEC indicated that most of the library members were forming soluble aggregates (Fig. 6). The aggregation can potentially reduce the accessibility of the histidine tag, decreasing the number of molecules bound to the resin, and can explain the low purification yield. Soluble aggregates represent almost the total amount of the purified protein for 2_1, 2_7, 2_9, 2_10, 2_12. The proteins 1_4, 2_4, 2_14 and the small peak of 2_7 elute at the same volume as YM_3A, indicating that they probably possess the correct shape of the consensus protein. 1_3 and 2_8 elute at earlier volume, indicating oligomerization, elongated shape or high flexibility.

The CD spectra are remarkably similar among all the proteins and YM_3A. The mean residue ellipticity is lower than in the consensus sequence when soluble aggregates are included in the CD sample. In the cases where it was possible to purify a monomer (1_4, 2_4,

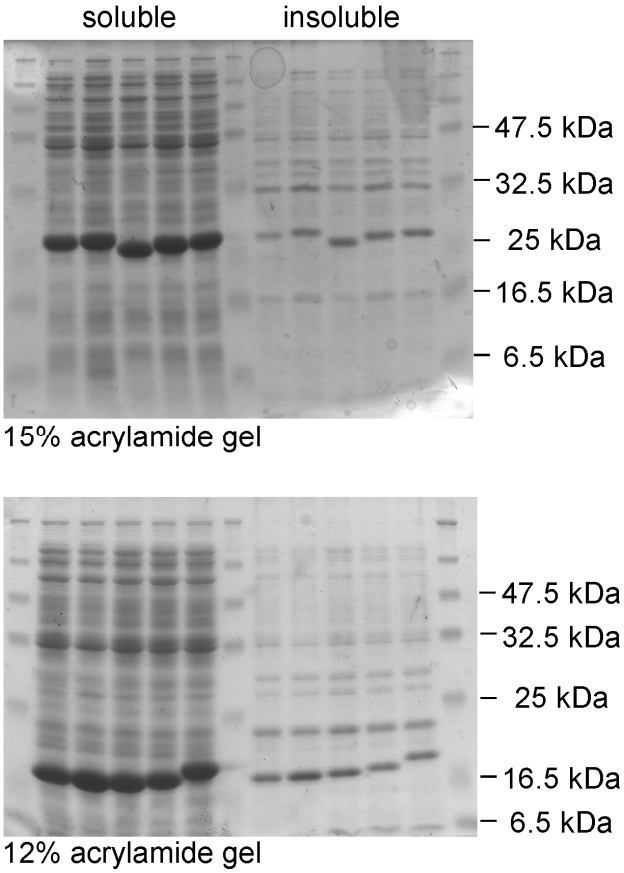


Fig. 5 Expression of unselected library members. In the upper gel are shown the soluble and insoluble fractions of, from left to right, 1_3, 1_4, 2_1, 2_4, 2_7. In the lower gel the other five clones 2_8, 2_9, 2_10, 2_12, 2_14 are shown. The expected size is approximately 23 kDa.

2_8, 2_14) the values were closer to the value of YM_3A (Fig. 7).

Many unselected members show strong ANS binding (Fig. 8). It is, however, difficult to discriminate if this behavior is the consequence of aggregation, of a hydrophobic randomized surface or of a molten globule-like state. ANS binding assay using the purified monomeric fraction of 1_4, 2_4, 2_8 and 2_14 led to approximately the same values for these proteins, indicating that the aggregation is probably not the reason, at least in these cases.

The stability of the protein was assessed by SDS-PAGE and mass spectrometry after 1 month at 4°C. In Ni-NTA elution buffer, containing Tris 50 mM, NaCl 0.5 M, glycerol 10%, imidazole 250 mM (pH 8), no sign of degradation was detected except for 1_3. Instead, degradation was observed after storage in Tris 20 mM, NaCl 50 mM (pH 8) except for 1_4, 2_4 and 2_10. The cleavage happened specifically in the N-cap after basic residues, as determined by the size of the fragments. The cause was probably a co-purified trypsin-like protease, which was not active in the first buffer tested. The results, summarized in Table 1, indicate that the more flexible and accessible part of the protein is the N-cap, but interaction with the internal repeats, and with randomized positions, can increase the stability as observed for 1_4, 2_4 and 2_10.

Table 1 Properties of unselected library members

	FT	Agg. SEC	No peak	ANS	Instability	Aromatics	Aliphatics
YM_3A					n.d.	3	3
1_3				x	x	1	6
1_4	x	x		x		2	4
2_1	x	x	x	x	x	4	5
2_4						2	3
2_7	x	x	x		x	3	3
2_8		x			x	5	2
2_9	x	x	x		x	5	4
2_10	x	x	x			5	4
2_12	x	x	x		x	3	3
2_14		x		x	x	1	2

FT: presence of protein in the flow through
Agg. SEC: presence of soluble aggregates in SEC
No peak: lack of monomeric (or clear oligomeric) peak in SEC
ANS: Strong ANS binding, above values of type-C consensus proteins
Instability: degradation products as detected by mass spectrometry (n.d. not determined)
Aromatics (F, W, Y): number of aromatic residues at the randomized positions
Aliphatics (I, L, M, V): number of aliphatic residues at the randomized positions

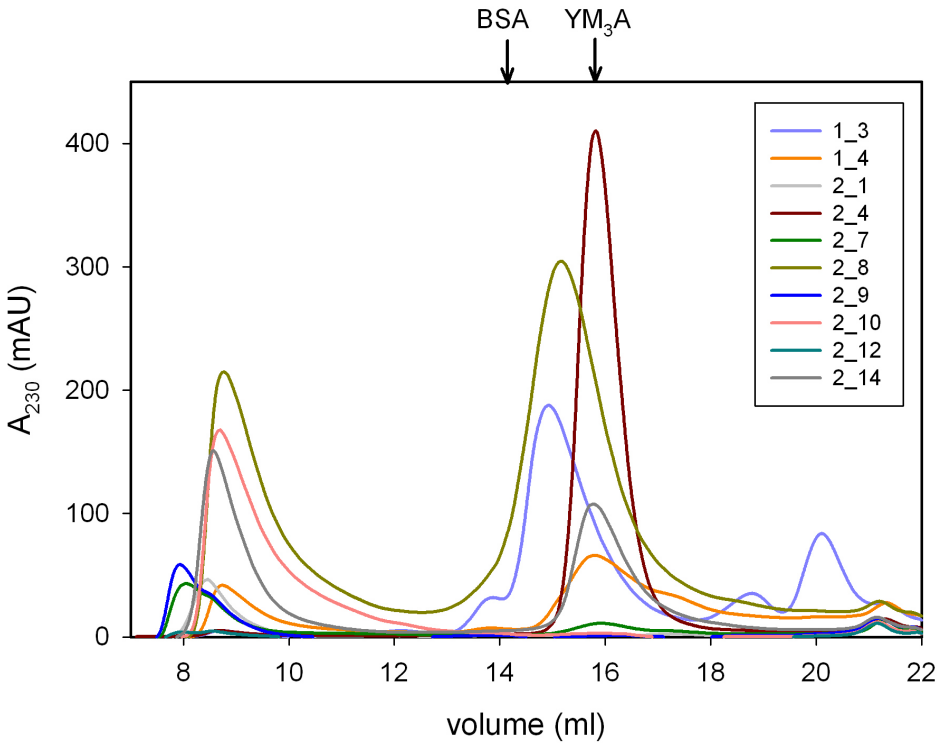


Fig. 6 Size exclusion chromatography of unselected library members. The elution volumes of BSA (66 kDa) and YM_3A (23 kDa) are used as reference and indicated by the arrows. The experiment was performed in 20 mM Tris, 50 mM NaCl, pH 8, with a Superdex 200 column. The elution was followed by absorbance at 230 nm; some of the unselected members lack tryptophan and do not show absorption at 280 nm.

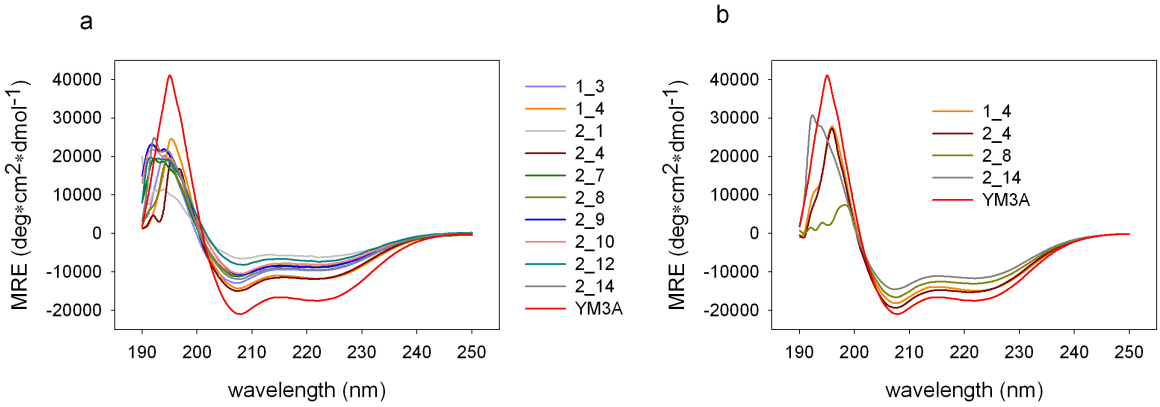


Fig. 7 CD spectra of unselected library members, containing soluble aggregates (a) or only monomers (b).

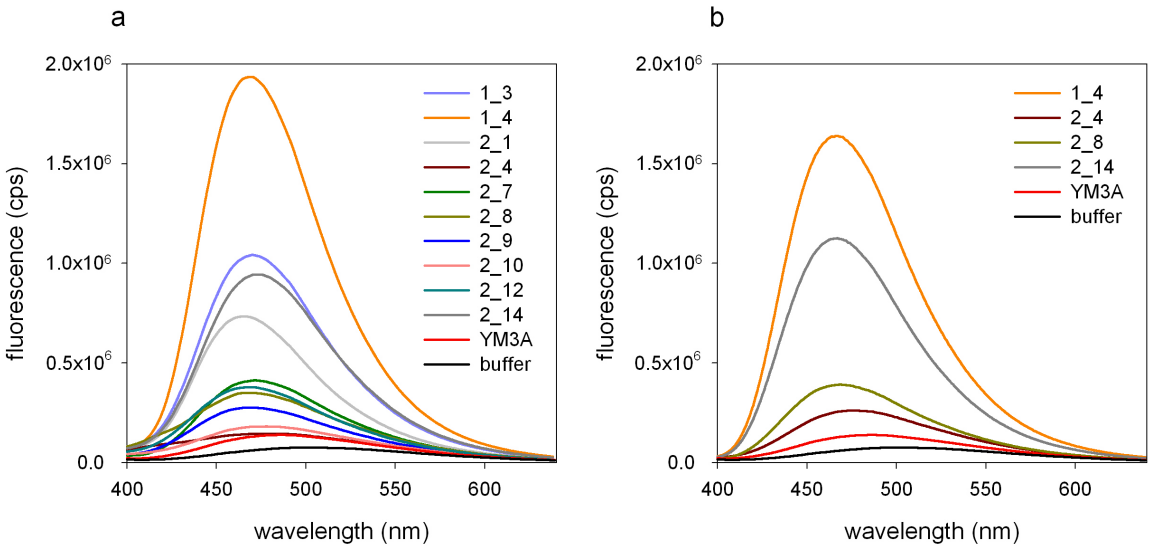


Fig. 8 ANS binding of unselected library members. The assay was performed with samples containing the soluble aggregates (a) or monomers (b). No significant difference was observed among the sample present in both experiments. Buffer is not subtracted but included as reference (in black) together with YM_3A (in red).

An analysis of the randomized surface could potentially provide hints to understand the tendency to aggregation and the reason for superior stability of certain library members. No crystal structure of any designed armadillo repeat protein is available and the interpretation is based on models. Unfortunately, no clear correlation was found between the observed aggregation propensity and the content of hydrophobic amino acids, even if, in general, an elevated number of exposed hydrophobic residues is detrimental for protein solubility. Certain library members possess several hydrophobic residues and aggregate (2_1, 2_9, 2_10) while others rich in aromatics (2_8) do not aggregate. If these hydrophobic residues are responsible for the aggregating behavior, the way this effect is modulated by their positions in the

structure and the neighboring residues is still elusive.

The analysis of unselect members indicated that is possible to obtain molecules, like 2_4, with CD spectrum, SEC profile and ANS affinity similar to the original consensus sequence. However, the majority of the molecules found have problems of aggregation, possibly leading to a disruption of the structure. A strong representation of hydrophobic residues on the surface of the proteins could be the reason for the aggregating behavior and the sensitivity to ANS binding.

A high percentage of molecules with these properties in the library could potentially represent a critical issue during selection, leading to selection of proteins able to form multimers or taking advantage of their hydrophobic surface to interact unspecifically with the target proteins.

Use of modified libraries based on KQ, QK and QQ mutants or construction of more stable libraries by insertion of additional consensus modules can improve the biophysical characteristics of the library members. Further refinement of the library for ribosome display selection is therefore described in Chapter 4.

References

1. Binz, H. K., Amstutz, P. & Plückthun, A. (2005). Engineering novel binding proteins from nonimmunoglobulin domains. *Nat Biotechnol* **23**, 1257-68.
2. Hosse, R. J., Rothe, A. & Power, B. E. (2006). A new generation of protein display scaffolds for molecular recognition. *Protein Sci* **15**, 14-27.
3. Skerra, A. (2007). Alternative non-antibody scaffolds for molecular recognition. *Curr Opin Biotechnol* **18**, 295-304.
4. Schneider, S., Buchert, M., Georgiev, O., Catimel, B., Halford, M., Stacker, S. A., Baechi, T., Moelling, K. & Hovens, C. M. (1999). Mutagenesis and selection of PDZ domains that bind new protein targets. *Nat Biotechnol* **17**, 170-5.
5. Malabarba, M. G., Milia, E., Faretta, M., Zamponi, R., Pelicci, P. G. & Di Fiore, P. P. (2001). A repertoire library that allows the selection of synthetic SH2s with altered binding specificities. *Oncogene* **20**, 5186-94.
6. Bosley, A. D. & Ostermeier, M. (2005). Mathematical expressions useful in the construction, description and evaluation of protein libraries. *Biomol Eng* **22**, 57-61.
7. Denault, M. & Pelletier, J. N. (2007). Protein library design and screening: working out the probabilities. *Methods Mol Biol* **352**, 127-54.
8. Neylon, C. (2004). Chemical and biochemical strategies for the randomization of protein encoding DNA sequences: library construction methods for directed evolution. *Nucleic Acids Res* **32**, 1448-59.

9. Virnekas, B., Ge, L., Plückthun, A., Schneider, K. C., Wellnhofer, G. & Moroney, S. E. (1994). Trinucleotide phosphoramidites: ideal reagents for the synthesis of mixed oligonucleotides for random mutagenesis. *Nucleic Acids Res* **22**, 5600-7.
10. Bradley, L. H., Kleiner, R. E., Wang, A. F., Hecht, M. H. & Wood, D. W. (2005). An intein-based genetic selection allows the construction of a high-quality library of binary patterned de novo protein sequences. *Protein Eng Des Sel* **18**, 201-7.
11. Munoz, V. & Serrano, L. (1994). Elucidating the folding problem of helical peptides using empirical parameters. *Nat Struct Biol* **1**, 399-409.
12. Lacroix, E., Viguera, A. R. & Serrano, L. (1998). Elucidating the folding problem of alpha-helices: local motifs, long-range electrostatics, ionic-strength dependence and prediction of NMR parameters. *J Mol Biol* **284**, 173-91.
13. Ernst, A. (2006). Combinatorial approaches to the evolution of novel proteins, University of Zürich

Chapter 4
Towards Ribosome Display Selection

Ribosome display principles

Ribosome display is a well established *in vitro* selection method based on the coupling of phenotype (protein) with genotype (mRNA) via the ribosome used for the protein synthesis [1; 2; 3; 4]. A scheme of the selection procedure is shown in Fig.1.

The original library is inserted, via digestion with restriction enzymes and ligation, in a ribosome display vector (pRDV, GenBank accession number AY327136 or pRDVhis, Appendix 4). A PCR product using outer primers T7b and tolAk (see oligonucleotide list, appendix 3) provides the template for *in vitro* transcription. The resulting RNA contains sequences at 5' and 3' forming stabilizing secondary structures to increase the resistance to exonucleases and thus the half-life.

The purified RNA is translated *in vitro* using an *E. coli* extract containing ribosomes, an energy souce, tRNA and amino acids. The library is fused in frame at 3' with a spacer derived from the *E. coli* gene *tolA*. The protein spacer keeps the library products distant from the ribosome, reducing its steric hindrance and allowing a better interaction with the target molecule. Stop codons are not present at the end of the coding sequence, preventing the dissociation of the ribosome and maintaining the link between phenotype and genotype.

The ternary complex protein-ribosome-RNA is incubated in contact with the target. Unbound complexes are washed away. The bound ternary complexes are dissociated and the RNA recovered.

A library subset is obtained after reverse transcription and PCR, using inner primers library-specific (see oligonucleotide list) to amplify only the library region. The resulting molecules can be inserted in an expression vector for single clone analysis or used as starting point for a second selection round. In the latter case the spacer and the 5' and 3' stabilizing sequences are reintroduced by inserting the PCR product in the ribosome display vector. The PCR product of the ligated plasmid will be the template for the next round.

Choice and format of the targets

A library containing three randomized internal modules can potentially bind 6-residue peptides. As initial target SV40 large-T antigen NLS sequence (KKKRKV) is an ideal

candidate. The interaction with importin α is well known and has been described in terms of affinity and kinetics [6]; therefore the properties of the selected binders can be directly compared.

Ribosome display targets are usually proteins or peptides expressed as fusion proteins (e.g. to λ phage protein D). The fusion protein contains peptide tags for purification, detection and immobilization. A histidine tag for purification and a biotinylated Avitag for immobilization are present (plasmid pAT223, Appendix 4). A linker region keeps the target peptide distant from the core of the fusion partner, preventing steric hindrance.

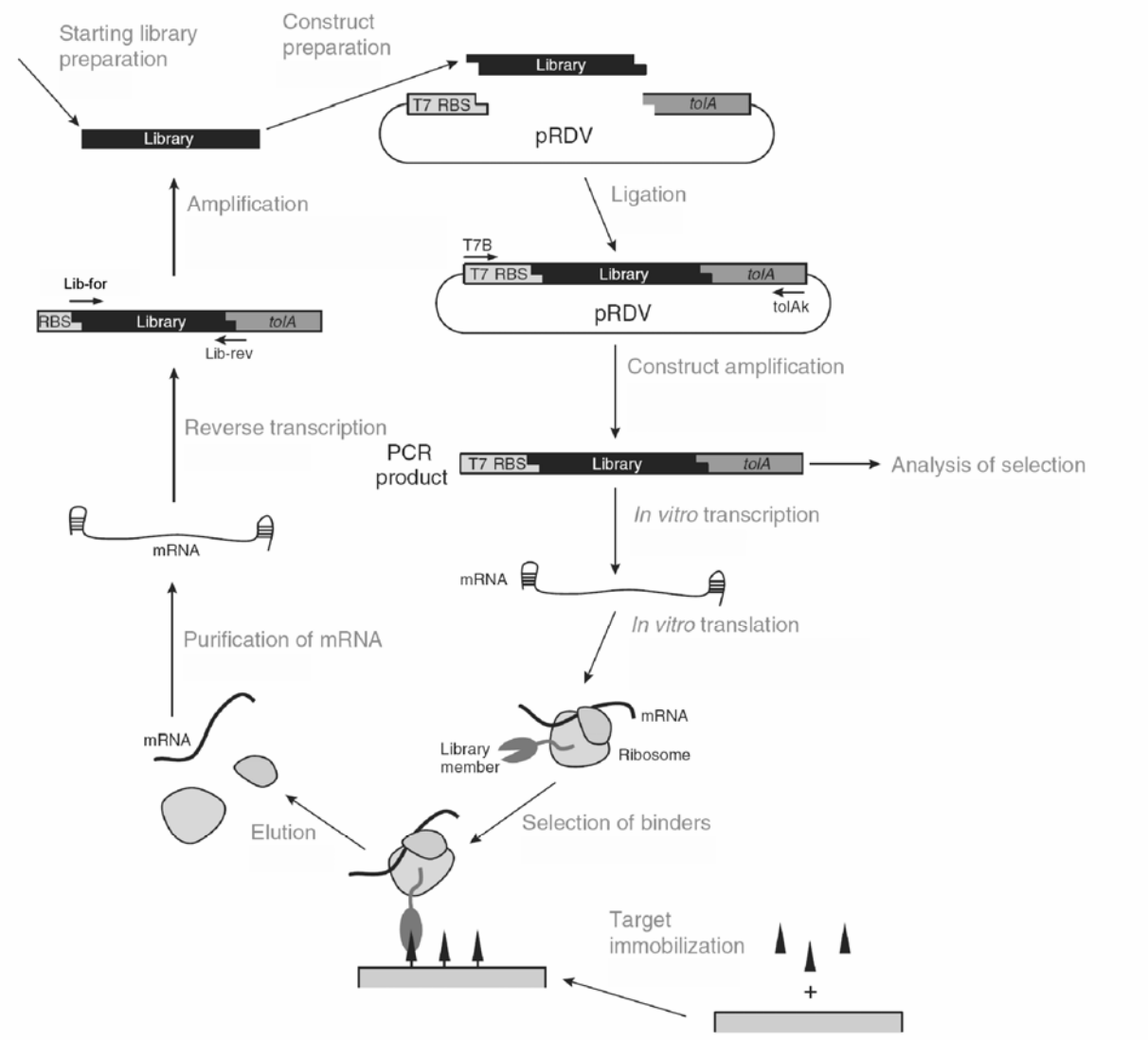
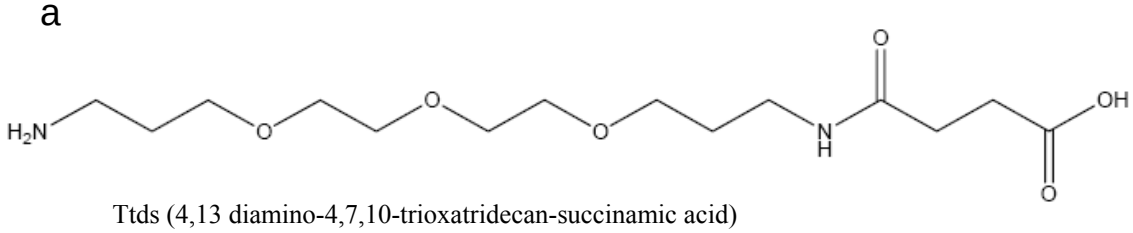


Fig. 1 Ribosome display selection scheme. Adapted from Zahnd et al. [4].

To avoid selections toward the fusion protein and especially the linker region, the target peptide was chemically synthesized by JPT (Berlin, Germany). The non-peptidic linker Ttds (1,13-diamino-4,7,10-trioxatridecan-succinamic acid) (Fig. 2a) was introduced between the target peptide and a biotin moiety required for binding to a neutravidin or streptavidin coated surface during the selection. Binding of the target peptide by importin α indicates that the peptide is accessible in the conditions used in ribosome display (Fig. 2b).

KKHTKK, KKYQKK and KKLDKK were chosen as additional target sequences in the same format. The selection of N3C binder could provide already information about residues in the first and third repeat employed to bind lysines.

a



Ttds (4,13 diamino-4,7,10-trioxatridecan-succinamic acid)

b

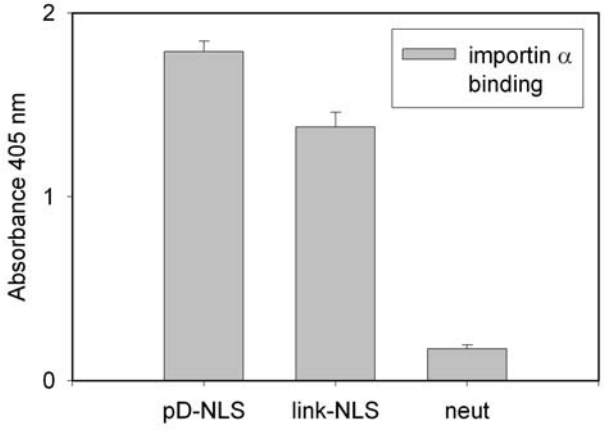


Fig. 2 Target format for ribosome display selection. (a) Structure of the non peptidic-linker Ttds, 4,13 diamino-4,7,10-trioxatridecan-succinamic acid.
The free N-terminus of the peptide was connected to the carboxy group of Ttds via an amide bond. Another amide bond was used to connect the carboxy group of biotin to the free amine group of Ttds. The carboxy group at the C-terminus of the peptide was modified into an amide group to avoid a selection bias toward free C-termini and to mimic a peptide bond. (c) ELISA of importin α binding to sv40 NLS sequence presented as fusion protein (pD-NLS) or with Ttds as non-peptidic linker (link-NLS). The fusion protein and the synthesized peptide were biotinylated and the ELISA was performed on plates coated with neutravidin, here shown alone as negative control (neut).

In vitro transcription and mRNA stability

The yield of *in vitro* transcription and the integrity of RNA are critical parameters for ribosome display selections. The RNA concentration can be estimated from the absorbance at 260 nm and the integrity verified on formaldehyde gels [7] (Fig. 3). Purifications using ethanol precipitation or spin columns (G50, GE Healthcare) were found to be almost equivalent, both in terms of yield and quality. No sign of degradation was visible in samples or controls. Treatment with DNAse or with the RNAse inhibitor vanadyl ribonucleoside complex did not affect the RNA quality. The weak smear in both armadillo and ankyrin libraries is probably due to a fraction of shorter molecules, always present and amplified during PCR cycles.

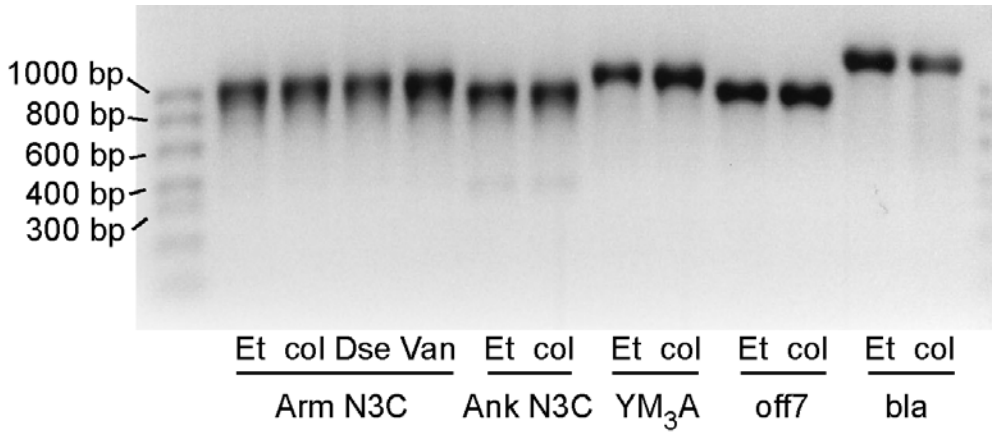


Fig. 3 RNA from *in vitro* transcription. Et and col indicate that the sample was purified using ethanol precipitation or size exclusion columns respectively; Dse, the sample was treated with DNAse; Van, vanadyl ribonucleoside complex was added as RNAse inhibitor. Arm and Ank N3C are armadillo and ankyrin libraries containing three internal modules, respectively, YM_3A is the consensus designed armadillo protein with three internal modules, off7 [8] is a selected ankyrin repeat protein with three internal modules, bla is β-lactamase. The expected size of RNAs are approximately 1100 nt for armadillo repeat proteins, 950 nt for ankyrin repeat proteins and 1300 nt for β-lactamase. On the left is indicated the RNA marker.

Preliminary selection results

The first results have been achieved using an N3C library based on the M-module and KKHTKK as target. After 5 selection rounds, several binders with low unspecific binding were detected in ELISA using the supernatant of lysed cells (data from G. Varadamsetty) (Fig. 4). Unfortunately, all the characterized binders were aggregation prone, as observed for

several unselected library members (Chapter 4). Therefore, the preliminary results indicate that binders can be selected from an armadillo repeat protein library.

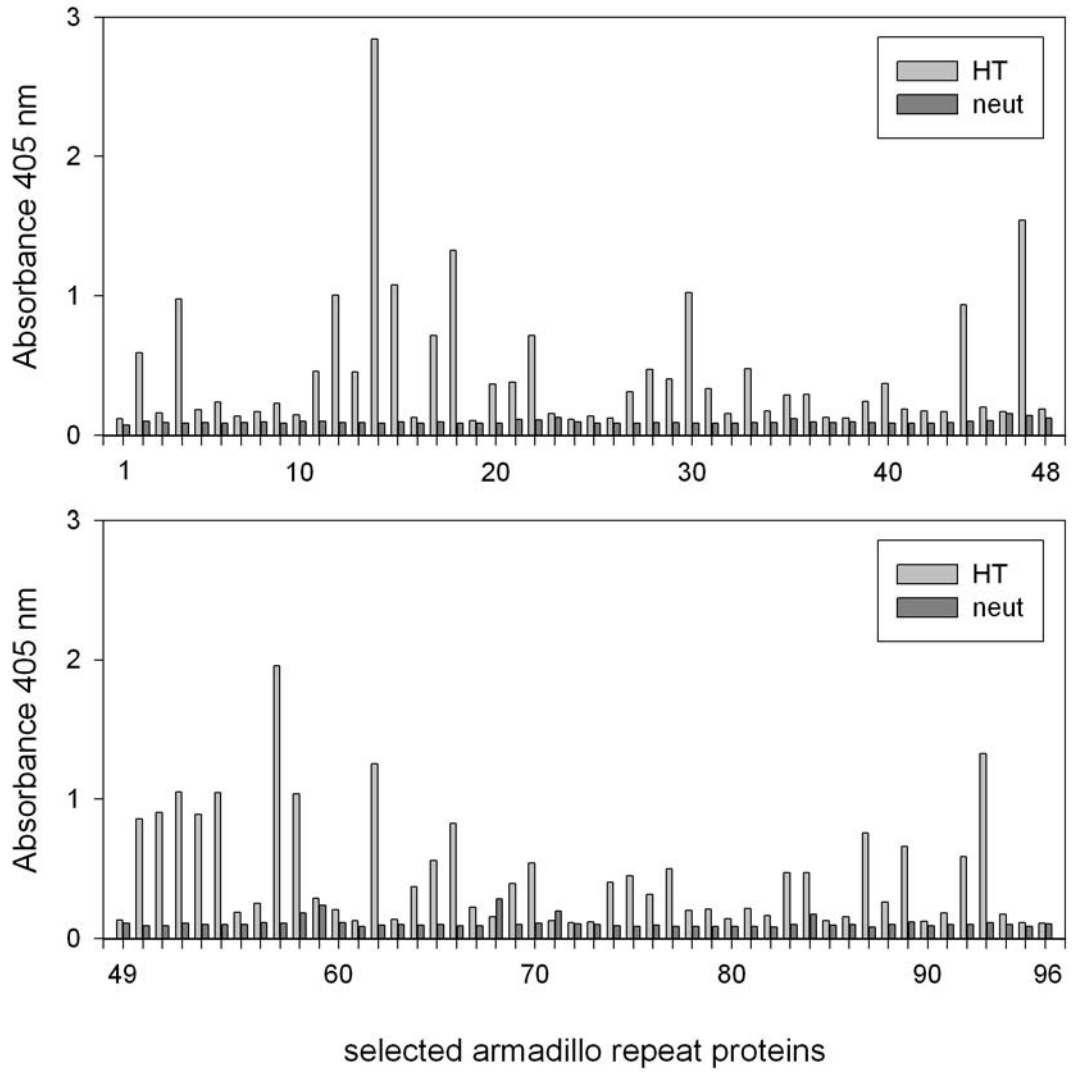


Fig. 4 Target recognition by selected armadillo repeat proteins. The ELISA of 96 single clones was carried out as described previously [5] and developed for 45 min. HT and neut indicate the target peptide KKHTKK, bound to neutravidin via biotin, and neutravidin alone, respectively.

Improvement of *in vitro* translation

The efficiency of *in vitro* translation in the ribosome display format was evaluated by radioactive translation of YM_4A, YM_3A and the armadillo library in comparison with β-lactamase, the designed ankyrin repeat protein library and the selected ankyrin off7 [8] (Fig. 5). Off7 shows an extraordinary translation efficiency and β-lactamase represents an average

reference point, while efficiency was noticed to be rather poor for armadillo repeat proteins (both for library and consensus). An increase in the amount of translated protein would thus positively influence the whole selection procedure.

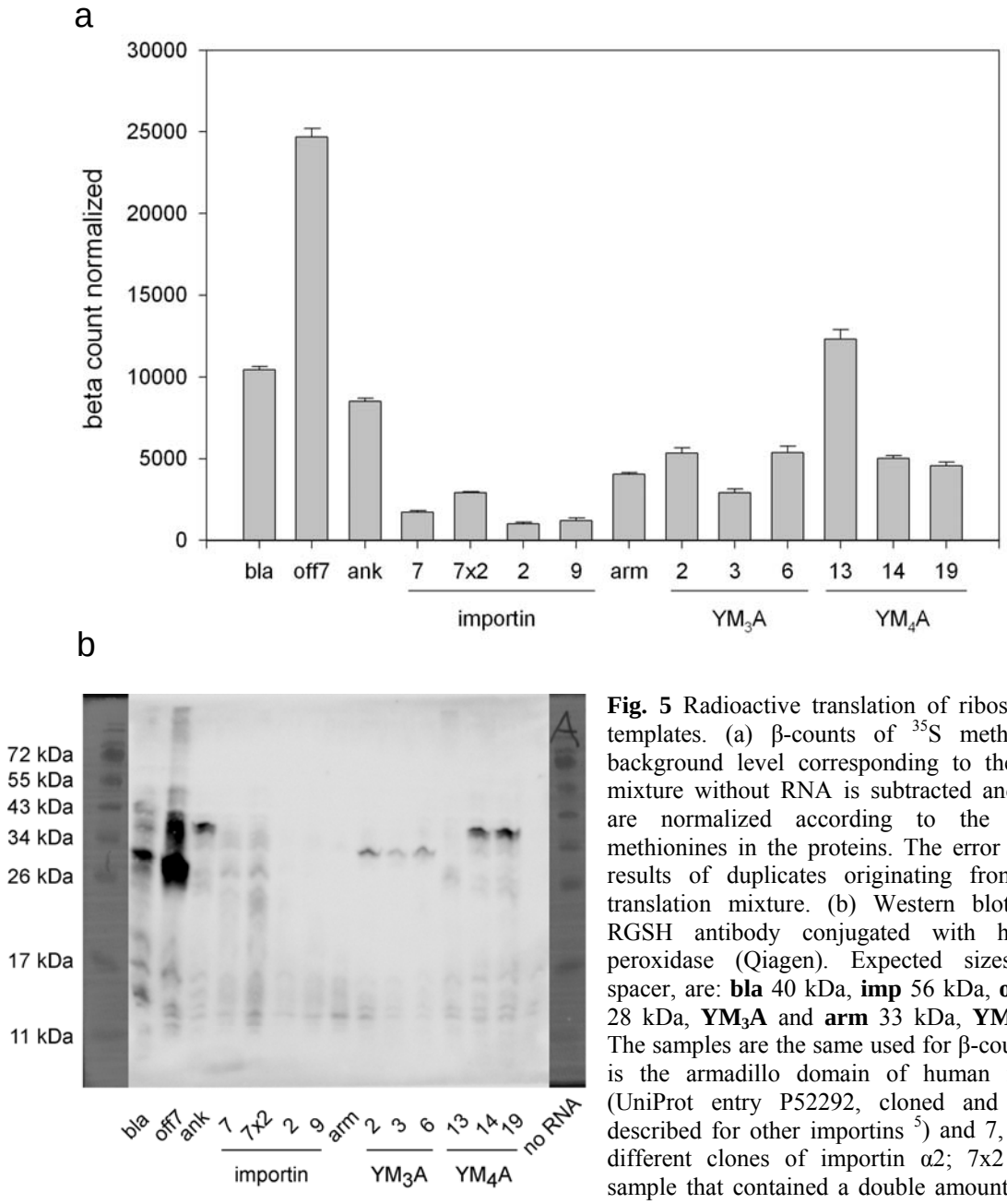


Fig. 5 Radioactive translation of ribosome display templates. (a) β-counts of ^{35}S methionine. The background level corresponding to the translation mixture without RNA is subtracted and the values are normalized according to the number of methionines in the proteins. The error bars are the results of duplicates originating from the same translation mixture. (b) Western blot with anti-RGSH antibody conjugated with horse radish peroxidase (Qiagen). Expected sizes, including spacer, are: **bla** 40 kDa, **imp** 56 kDa, **off7** and **ank** 28 kDa, **YM_3A** and **arm** 33 kDa, **YM_4A** 36 kDa. The samples are the same used for β-count. Importin is the armadillo domain of human importin α2 (UniProt entry P52292, cloned and purified as described for other importins [5]) and 7, 2 and 9 are different clones of importin α2; 7x2 indicates a sample that contained a double amount of template RNA from clone 7. 2, 3 and 6 are different clones of YM_3A. 13, 14 and 19 are different clones of YM_4A. Ank and Arm are N3C ankyrin and armadillo library, respectively. Off7 is a selected ankyrin repeat protein, bla indicates β lactamase. The clones possess the same sequences. Variations are probably due to effects dependent on the single translation experiments.

In the absence of a limiting factor for the translation (e.g. elongation factor, energy supply, tRNA and amino acids) a low protein production could be dependent on mRNA degradation, protein degradation or problems in translation initiation.

As shown above, the starting RNA is not degraded and the proteins produced are not particularly susceptible to protein degradation, as indicated by the level of expression achieved and the absence of degradation products.

Initiation of translation is strongly related to the availability of the Shine-Dalgarno (SD) sequence and the initial AUG codon [9]. These sequences can be involved in the formation of RNA secondary structure that can limit the ribosome accessibility. RNA secondary structures can be disrupted or prevented *in vivo* by RNA binding proteins, but the absence of these factors *in vitro* can potentially hamper the initiation of translation.

A version of the program M-fold [10; 11; 12], implemented in GCG Package (Version 11.1.2, Accelrys Inc., San Diego, CA), was used to calculate possible RNA secondary structures able to sequester the SD sequence or the initial AUG codon.

The circle representations of RNA structures indicate that the SD sequence is free for interaction with the ribosome in YM_3A, while the position 110 (the third base of the initial AUG) is involved in the formation of a stem loop, as well as the following bases (Fig. 6a). SD and AUG of off7 and β-lactamase are involved only in the formation of small hairpins and short-range interactions. The formation of a stable secondary structure could thus be responsible for the poor translation rate. The stem loop formed by YM_3A occurs in all the calculated structures and involves always the sequence coding for the $MRGSH_6$ tag and a sequence present in the framework of the M-module (corresponding to the positions 28-31). In fact, the calculated secondary structures show formation of the stem loop with all the three internal modules.

Silent mutations in both in the N-terminal tag and in the framework positions were tested *in silico* for their ability to disrupt the stem loop formation (Fig. 6a). The mutations were identified with a three-letter name corresponding to the bases present at the three positions susceptible of change (Fig. 6b). Among the possible versions, CAG and CCC mutations were introduced in the pRDVhis vector (Appendix 4), containing YM_3A or β-lactamase, by site directed mutagenesis (Stratagene Quickchange protocol). The oligonucleotides are described in Appendix 3.

a

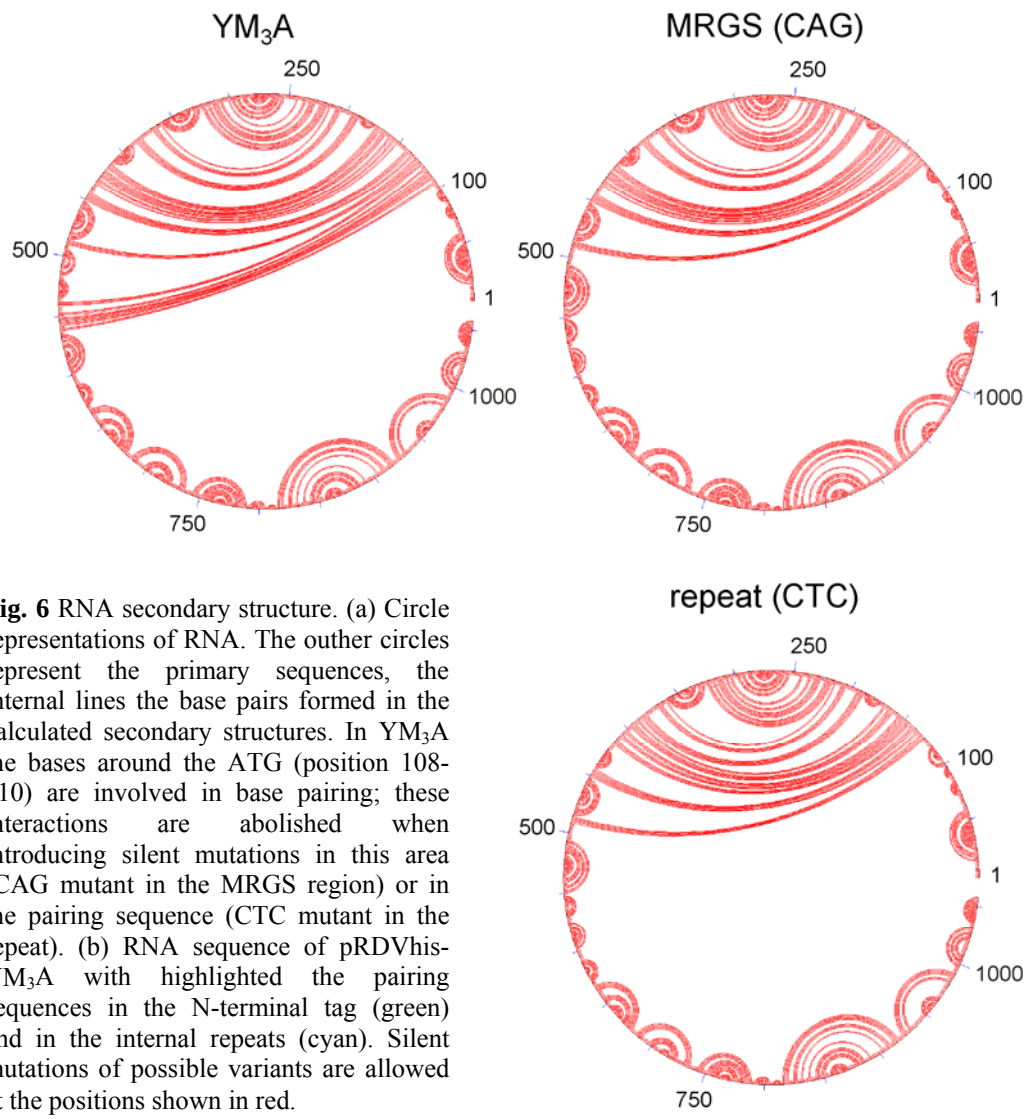


Fig. 6 RNA secondary structure. (a) Circle representations of RNA. The outher circles represent the primary sequences, the internal lines the base pairs formed in the calculated secondary structures. In YM_3A the bases around the ATG (position 108-110) are involved in base pairing; these interactions are abolished when introducing silent mutations in this area (CAG mutant in the MRGS region) or in the pairing sequence (CTC mutant in the repeat). (b) RNA sequence of pRDVhis-YM_3A with highlighted the pairing sequences in the N-terminal tag (green) and in the internal repeats (cyan). Silent mutations of possible variants are allowed at the positions shown in red.

b

pRDVhis21-YM_3A

AUACGAAAUUAAUACGACUCACUAUAGGGAGACCACAACGGUUUCCCUAAUUGUGAGCGGAUAACAAUAGAAAUA
M R G
AUUUUGUUUAACUUUAAGAAGGAGAUAUAUUCAUGAGAGAGGAUCGCAUCACCAUCACCAUCACGGAUCCGAACUGC
CGCAGAUGACCCAGCAGCUGAACUCUGACGACAUGCAGGAACAGCUGUCUGCUACCGUUAAAUUCCGUCAGAUCC
UGUCUCGUGAUGGU
AACGAACAAAUCCAAGCUGUUAUCGAUGCUGGUGCUCUGCCGGCUCUGGUUCAACUGCUGUCCUCUCCGAACGAG
AAGAUCCUGAAAGAAGCUCUGUGGGCUCUGUCUAACAUCGCUUCUGGUGGU
AACGAACAAAUCCAAGCUGUUAUCGAUGCUGGUGCUCUGCCGGCUCUGGUUCAACUGCUGUCCUCUCCGAACGAG
AAGAUCCUGAAAGAAGCUCUGUGGGCUCUGUCUAACAUCGCUUCUGGUGGU
AACGAACAAAUCCAAGCUGUUAUCGAUGCUGGUGCUCUGCCGGCUCUGGUUCAACUGCUGUCCUCUCCGAACGAG
AAGAUCCUGAAAGAAGCUCUGUGGGCUCUGUCUAACAUCGCUUCUGGUGGU
AACGAACAGAAACAGGCUGUUAAAGAAGCUGGUGCUCUGGAGAAACUGGAACAGCUGCAGUCCCACGAGAACGAG
AAGAUCCAGAAAGAAGCUCAGGAAGCUCUGGAGAAGCAGUUCUCCCAC
AAGCUUUAUAUGGCCUCGGGGGCCGAAUUCGGAUCUGGUGGCCAGAAGCAAGCUGAAGAGGCGGCAGCGAAAGCG
GCGGCAGAUGCUAAAGCGAAGGCCGAAGCAGAUGCUAAAGCUGCGGAAGAAGCAGCGAAGAAAGCGGCUGCAGAC
GCAAAGAAAAAAGCAGAAGCAGAAGCCGCCAAAGCCGCAGCCGAAGCGCAGAAAAAAGCCGAGGCAGCCGCUGCG
GCACUGAAGAAGAAAGCGGAAGCGGCAGAAGCAGCUGCAGCUGAAGCAAGAAAGAAAGCGGCAACUGAAACCGCA
CACCUUACUGGUGUGCGG

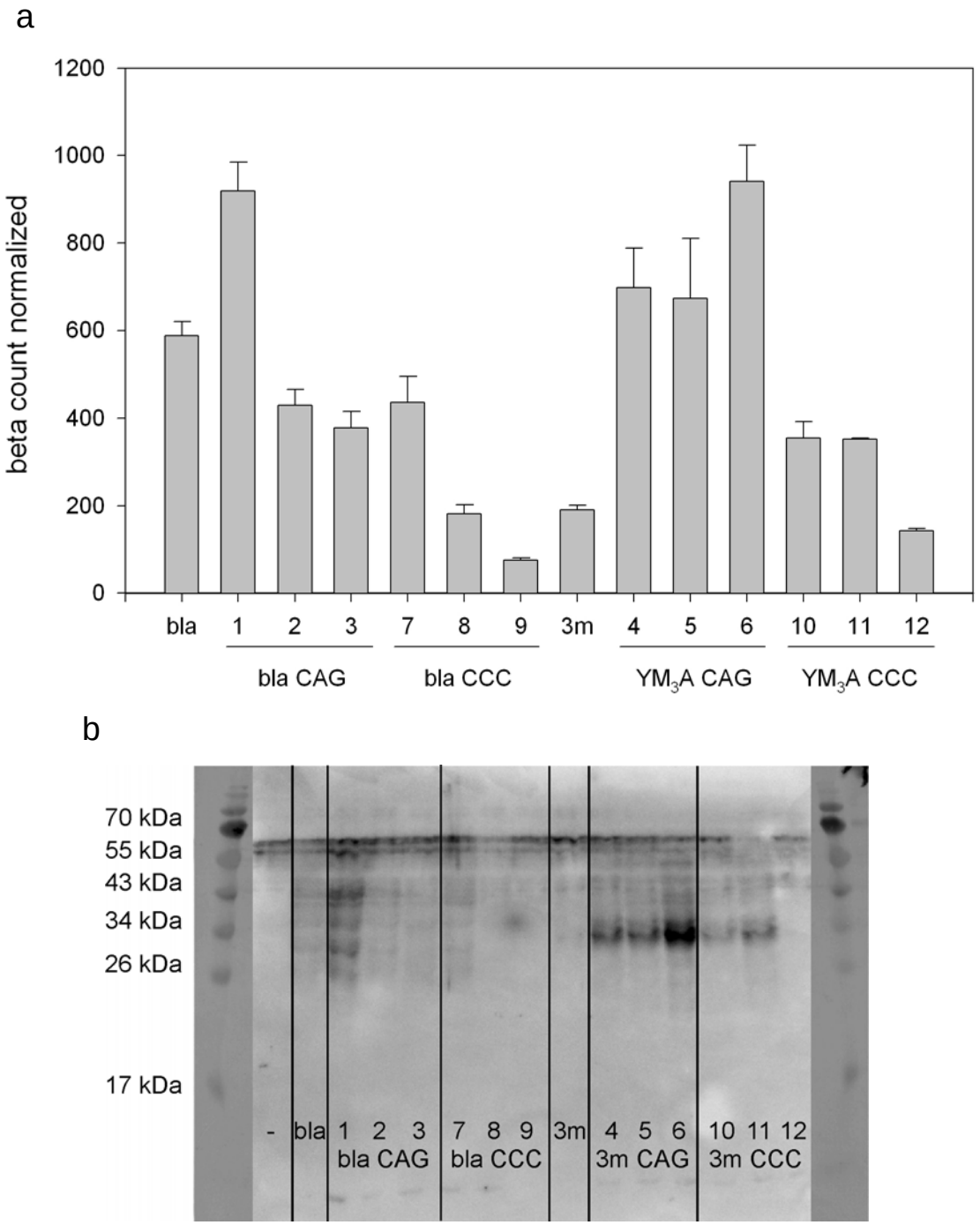


Fig. 7 *In vitro* translation after RNA secondary structure disruption. (a) β-counts of ^{35}S methionine. The background level corresponding to the translation mixture without RNA is subtracted and the values are normalized according to the number of methionines in the proteins. The error bars are the results of duplicates originating from the same translation mixtures. (b) Western blot with anti RGSH antibody conjugated with horse radish peroxidase (Qiagen). Expected sizes, including spacer, are: **bla** 40 kDa, **YM_3A** 33 kDa. The samples are the same used for β-counting; **bla** indicates β-lactamase, **3m** is YM_3A and – refers to a negative control without RNA. 1, 2, 3 are clones containing β lactamase and CAG mutation. 4, 5, 6 are clones containing YM_3A and CAG mutations. 7, 8, 9 are clones containing β lactamase and CCC mutations. 10, 11, 12 are clones containing YM_3A and CCC mutations. The clones possess the same sequences. Variations are probably due to effects dependent on the single translation experiments.

The CAG version of YM_3A led to a threefold increase in translation efficiency (Fig. 7), reaching a slightly higher level than the reference β-lactamase, and confirming the original hypothesis.

The vector pRDVhis-CAG (see Appendix 4) containing the silent mutations will thus be used for the next rounds of selection of armadillo repeat proteins.

N5C libraries

New libraries containing five internal modules based on KK, KQ and QQ modules were generated as described in Chapter 3, with a complexity of approximately $5x10^{11}$. The names indicate the presence of lysine or glutamine at position 26 or 29 in each repeat. The type-M module contains lysine at both position and correspond thus to KK (see chapter 2). The libraries are formed by an N-terminal Ny capping repeat, a module not randomized, three randomized modules, one module not randomized and the C-terminal Ca capping repeat. All the internal modules are of the same type. The new libraries were based on the observation (by G. Varadamsetty) that, when introducing two modules not randomized on the sides of the randomized ones, the previously aggregating proteins were becoming monomeric in SEC and showing all the characteristics typical of M-type proteins. The same result was observed for the unselected library members. The new N5C libraries will be used for the next rounds of selection of armadillo repeat proteins.

Materials and methods

Radioactive translation. Radioactive translation was performed for 10 min at 37 °C according to the published ribosome display protocol [4], using 5 µl of 10 µM ^{35}S-methionine in a final volume of 12.5 µl. 5 µl of the sample were used for gel preparation and 7.5 µl for radioactivity measurement. After addition of 9 volumes (67.5 µl) of 0.6 M potassium hydroxide the samples were incubated at 37 °C for 5 min to promote the hydrolysis of the bond between tRNA and protein or amino acids, without significantly degrading the translated protein.

Proteins were precipitated by adding 9 volumes (675 µl) of 25% trichloroacetic acid (TCA) containing 2% casein amino acids and incubating in ice for 30 min. Samples were filtrated through a 96-well filter plate (Multiscreen HTS FB, Millipore) preconditioned with 5% TCA. The filters were washed 4 times with 300 µl 5% TCA and then transferred to a new 96-well plate, where 175 µl of Optiphase supermix liquid scintillation cocktail (Perkin Elmer) were added. The samples were measured after 3 h incubation.

References

1. Hanes, J. & Pluckthun, A. (1997). In vitro selection and evolution of functional proteins by using ribosome display. *Proc Natl Acad Sci U S A* **94**, 4937-42.
2. Schaffitzel, C., Hanes, J., Jermutus, L. & Pluckthun, A. (1999). Ribosome display: an in vitro method for selection and evolution of antibodies from libraries. *J Immunol Methods* **231**, 119-35.
3. Schaffitzel, C., Zahnd, C., Amstutz, P., Luginbühl, B. & Plückthun, A. (2005). In Vitro Selection and Evolution of Protein-Ligand Interactions by Ribosome Display. In *Protein-protein interactions : a molecular cloning manual* 2nd ed. edit. (Golemis, E. & Adams, P. D., eds.), pp. xiv, 938 p. Cold Spring Harbor Laboratory Press, New York.
4. Zahnd, C., Amstutz, P. & Plückthun, A. (2007). Ribosome display: selecting and evolving proteins in vitro that specifically bind to a target. *Nat Methods* **4**, 269-79.
5. Parmeggiani, F., Pellarin, R., Larsen, A. P., Varadamsetty, G., Stumpp, M. T., Zerbe, O., Caflisch, A. & Pluckthun, A. (2008). Designed armadillo repeat proteins as general peptide-binding scaffolds: consensus design and computational optimization of the hydrophobic core. *J Mol Biol* **376**, 1282-304.
6. Catimel, B., Teh, T., Fontes, M. R., Jennings, I. G., Jans, D. A., Howlett, G. J., Nice, E. C. & Kobe, B. (2001). Biophysical characterization of interactions involving importin-alpha during nuclear import. *J Biol Chem* **276**, 34189-98.
7. Sambrook, J. & Russell, D. W. (2001). *Molecular cloning : a laboratory manual*. 3rd edit, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.
8. Binz, H. K., Amstutz, P., Kohl, A., Stumpp, M. T., Briand, C., Forrer, P., Grütter, M. G. & Plückthun, A. (2004). High-affinity binders selected from designed ankyrin repeat protein libraries. *Nat Biotechnol* **22**, 575-82.
9. Makrides, S. C. (1996). Strategies for achieving high-level expression of genes in Escherichia coli. *Microbiol Rev* **60**, 512-38.
10. Zuker, M. (1989). On finding all suboptimal foldings of an RNA molecule. *Science* **244**, 48-52.
11. Jaeger, J. A., Turner, D. H. & Zuker, M. (1989). Improved predictions of secondary structures for RNA. *Proc Natl Acad Sci U S A* **86**, 7706-10.
12. Mathews, D. H., Sabina, J., Zuker, M. & Turner, D. H. (1999). Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol* **288**, 911-40.

Chapter 5
Computational Design of an Armadillo Scaffold

A different approach

The characteristics of designed armadillo repeat proteins and the first results from libraries (see chapter 4) are promising, but improvements are still possible already at the level of the framework. The major issue comes from the analysis of binding modes in models of long armadillo repeat proteins. With the increase of the number of repeats for binding of longer peptides, the target backbone requires a significant distortion to be able to bind the armadillo repeat protein in the desired way. This could turn into a reduced affinity or specificity if the armadillo domain is not able to adapt its conformation. A design taking into account the binding geometry and the reciprocal orientation of the repeats, which is responsible for the binding geometry, will improve dramatically the possibility to achieve high affinity and specificity, reducing at the same time the entropy loss due to the adaptation of the armadillo protein to the required geometry.

Such a control of the atomic details cannot be achieved using consensus design but requires computational tools able to build a structure with the desired characteristics and select the appropriate sequence to realize it.

Design with Rosetta

Among the available software for computational design ROSETTA [1] arose in the last years as one of the prominent program for protein design, due to its open approach based on a wide spread community and the successes in the field of structure prediction and protein design. Originally developed as structure prediction tool [2], ROSETTA achieved recently some of the most interesting results ranging from the *de novo* design of a new folding [3] to the development of new enzymes [4; 5].

ROSETTA is a statistically-based design program that relies on assembly of small fragments extracted from the Protein Data Bank (PDB) and combined using a Monte Carlo simulated annealing approach. The different insertions are accepted or rejected based on Bayes theorem [2]. The underlying hypothesis of the fragment choice (3-9 residues long) is that the local sequence determines the local conformation.

The design of a new sequence based on a fixed backbone is composed of a first approximated search with a simplified potential, using for the amino acids only the backbone and a centroid atom centered on the Cβ of the side chains. The atomic details are completely missing, but the process easily discards the incompatible conformations, increasing considerably the speed of the whole process. A second refinement step is performed in the same way, using a complete force field and full atomic details of the amino acids [1].

Besides the continuous software development, this approach is becoming more and more powerful due to the increase of available structures and the consequent increasing richness of the fragment database.

A new armadillo framework

The first step in designing a new geometrically defined armadillo framework was to identify the desired orientation of adjacent repeats, in order to obtain a correctly positioned binding site without distortion of the target peptide.

The work of Annemarie Honegger, who collaborated on this part of the project, was focused on the analysis of repeat-repeat interactions and peptide binding and the definition of their geometrical parameters (Fig. 1). Based on this analysis, models of armadillo repeat proteins containing several repeats with the same parameters were built and evaluated by their interaction with a poly-alanine chain (Fig. 2).

The template with lowest distortion of the target peptide was chosen as backbone for redesign, introducing backbone flexibility to improve the search. The iterative process of design and evaluation was done in collaboration with Dr. Sarel Fleishman, in the Baker group at the University of Washington, Seattle, USA.

The hydrophobic core of the internal repeats was redesigned first: this is usually the most critical but also the most reliable part of the design, because it is dominated by van- der-Waals interactions than can be easily calculated with a Lennard-Jones potential.

ROSETTA provides always a set of structures that have to be evaluated, with energetic criteria or based on the experience of the user, and the most reasonable ones are allowed to proceed towards the next design steps.

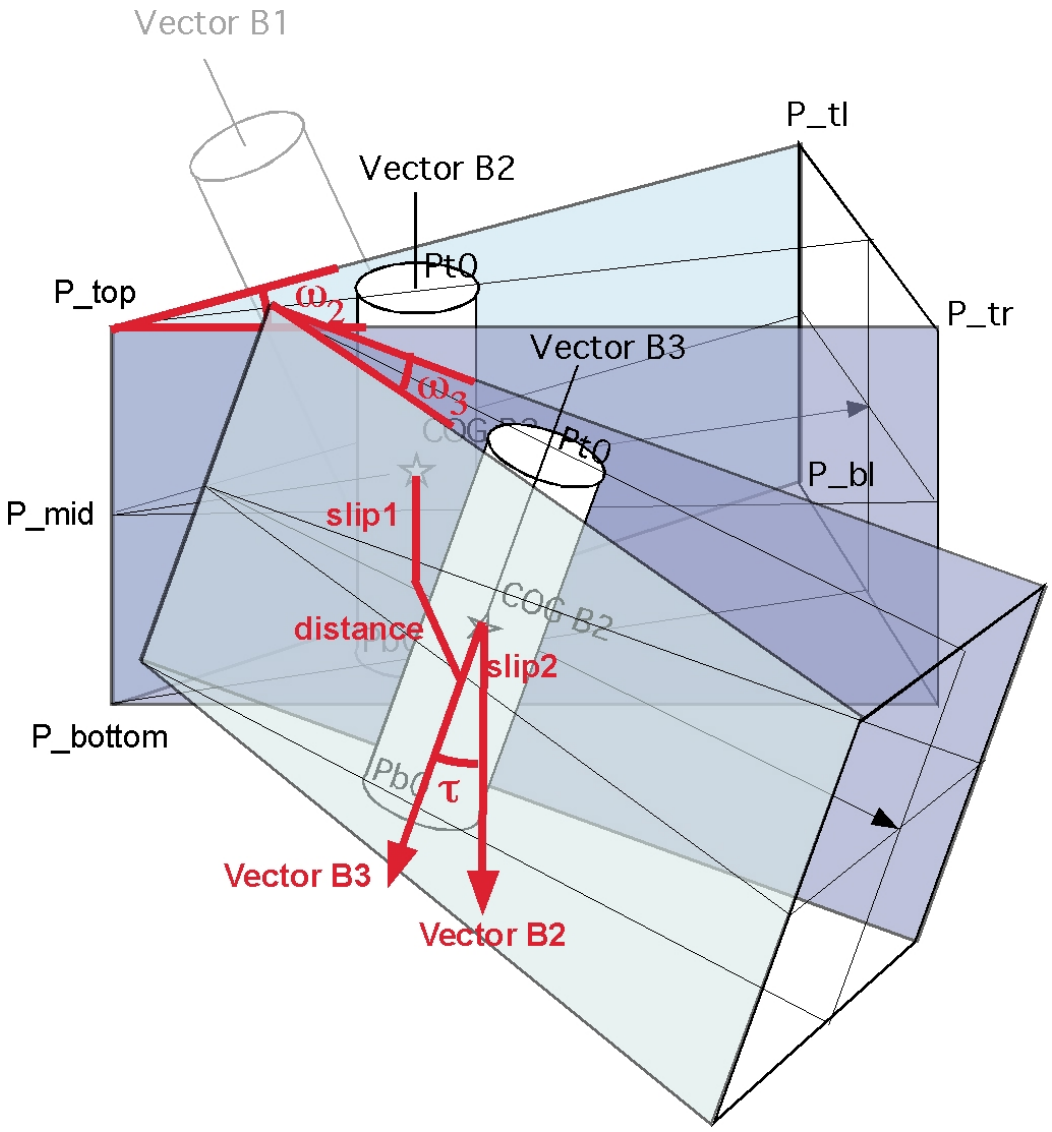


Fig. 1 Definition of relative orientation of armadillo repeats. Each repeat is considered a wedge shaped object whose geometry is determined by the relative orientation of the B-helices (H3 in the usual nomenclature) of the adjacent repeats.

The planes separating two adjacent repeats are parallel to the B-helix vectors of the two repeats and equidistant from the centers of gravity of the two B-helices. The angle between the two planes separating repeat n from repeat n-1 and from repeat n+1 is the wedge angle ω. For graphical representation, the wedge is represented bounded by top and bottom planes, perpendicular to the B-vector, and a back plane perpendicular to the plane bisecting the ω angle.

The parameters slip1 and slip2 describe the footpoints of the distance vector on the helix axes. They are defined as the distance between the Center of Gravity (CoG, indicated by a star) of a helix, which is located on the helix vector, and the footpoint of the distance vector. If slip is a positive number, it is measured from CoG in the direction of the helix vector, if it is negative, it is measured in the opposite direction. As defined, the midpoint of the distance lies on the plane separating the two repeats, and the distance vector is perpendicular to this plane. An additional rotation angle is needed to fully describe the relative orientation of the two repeats: the tilt angle τ around the distance vector, corresponding to the angle between vector B2, and vector B3. The picture and the description are modified versions of the original documentation from A. Honegger.

The second part involved the redesign of the surface of internal repeats. Several solutions were available, due to the lack of packing constraints, characteristic of the core, and the space available for side chains. Therefore, preferred residues were not always clear for the surface positions. The sensitivity of the energy function to electrostatics and polar terms is a common drawback in all design programs and force fields, due to the complexity of interactions involved and the approximated mathematical description necessary to allow calculations in a reasonable time. In ROSETTA this factor is strongly approximated using one term for formation of charged pairs and one term for solvent accessibility: the solution is often a series of possible alternatives that should be analyzed. Most of the choices at this stage were taken considering the most common residues appearing in the natural armadillo repeat proteins. During these two parts a symmetry constraint was applied to introduce the same variations in every repeat.

As third step the capping repeats were redesigned: the exposed hydrophobic residues were replaced by hydrophilic amino acids to protect the protein hydrophobic core. As for the internal repeat surface, several positions were not clearly defined and additional information, from natural proteins or from amino acid characteristics (like helical propensity), was used to choose the residues to be placed at these positions.

The final sequences after the design in collaboration with S. Fleishman are shown in Fig. 3. It can be noticed that one of the two lysines responsible for the pH effect described in chapter 2 has been replaced by a glutamate at position 26, introducing a stabilizing negative charge. The putative stability of a model containing two internal repeats and the new designed capping sequences was assessed by molecular dynamics at 300 K in implicit solvent. The result of the simulations performed by Dr. Enrico Guarnera in the group of Prof. A. Caflisch revealed instability of the helix 3, probably due to the presence of serines with unfavorable helical propensity at positions 33 and 36 (Fig. 4). Alanines were introduced in the sequence to replace the serines and the increase in stability was confirmed by simulations with the mutant version.

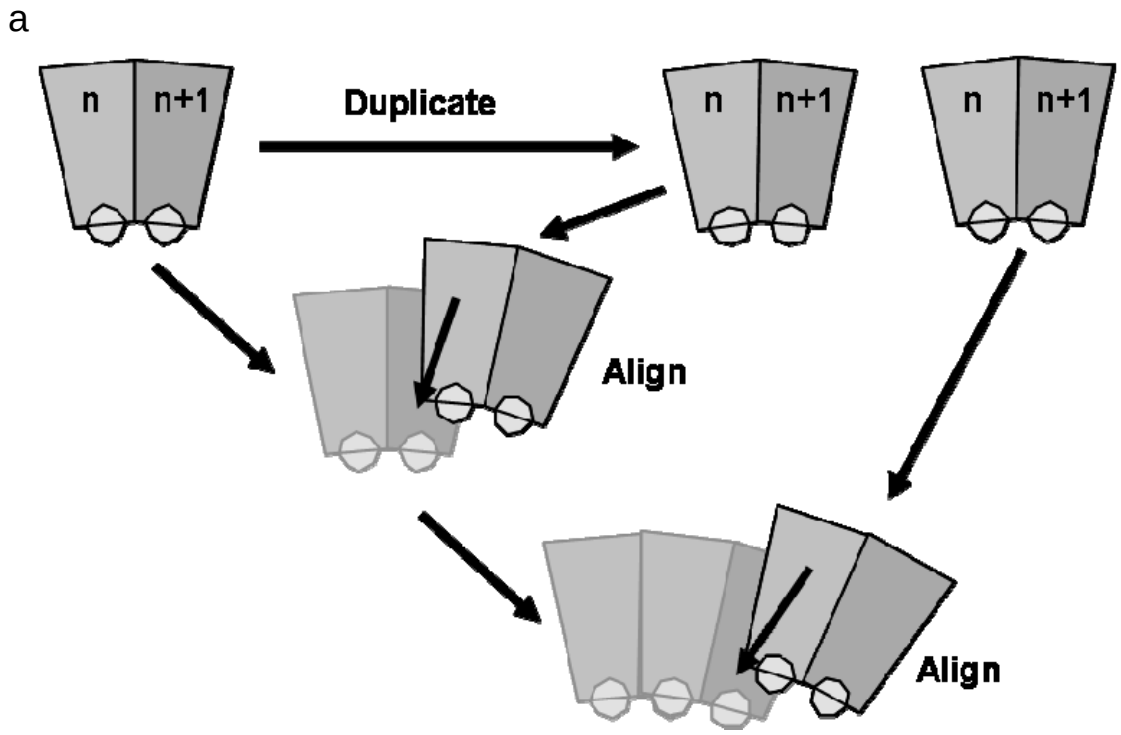
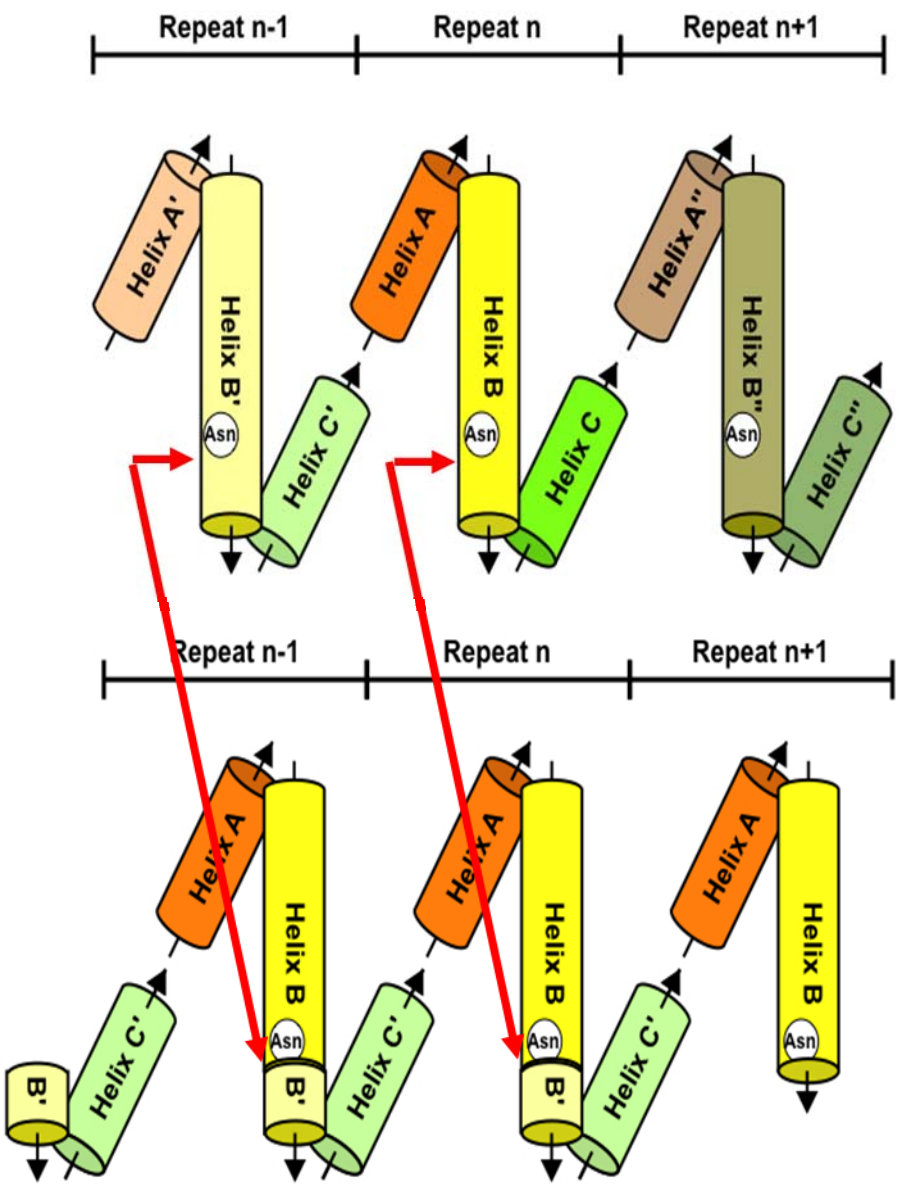
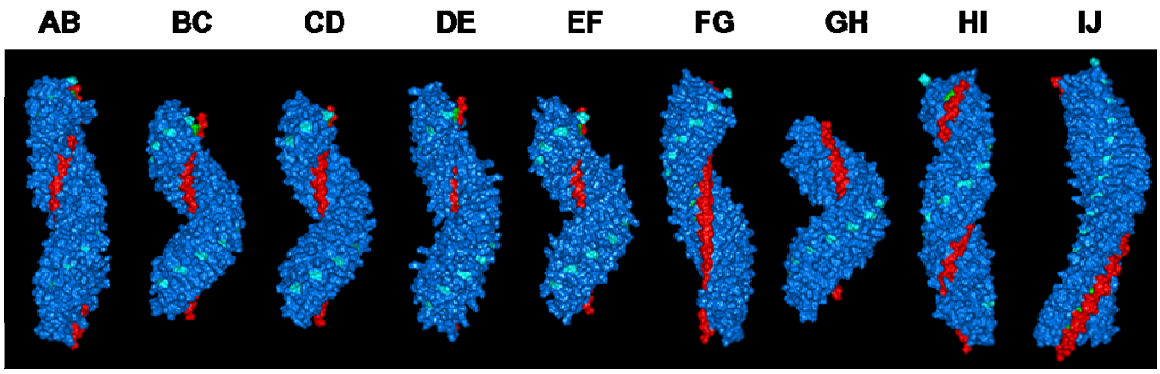


Fig. 2 Construction of multi-repeat models. (a) Schematic drawing of the construction procedure. A double repeat was generated from two repeat fragments and the multi-repeat was built by superimposing the first repeat of the second copy on the second repeat of the first copy, then deleting either the first or the second repeat of each copy and joining up the remaining residues. (b) Construction of single repeat. In this model the single repeat is centered on the helix3 (called here B) which is the main determinant for the binding site formation and the geometry. Because the interface has to be conserved, the single repeat to be used in the multi-repeat construction was generated from two different parts contributing to the formation of the interface. The junction point is after the conserved asparagine at position 37 (Asn in the picture). The part before is coming from the repeat n, the part after from the repeat n-1. The red arrows indicate the cuts and the corresponding junction points. (c) Multi-repeat models. Each model was built using the interface between two consecutive repeats, indicated by the letters, for the construction of the single repeat, and then used for the construction of the dimer including the geometrical information. A model corresponding to each couple of repeats from importin yeast, importin mouse and catenin was realized. The multi repeat models based on yeast importin are shown here. A polyalanine peptide bound to the model is depicted in red. The model GH was used as starting point for the sequence design, because of the most favorable geometry for peptide binding without distorsion of the target. The picture and the description are modified versions of the original documentation from A. Honegger.

b


Repeat n-1
Repeat n
Repeat n+1
Helix A'
Helix B'
Helix C'
Helix A
Helix B
Helix C
Helix A''
Helix B''
Helix C''
Asn
Repeat n-1
Repeat n
Repeat n+1
Helix A
Helix B
Helix C'
B'
Asn

c


AB
BC
CD
DE
EF
FG
GH
HI
IJ

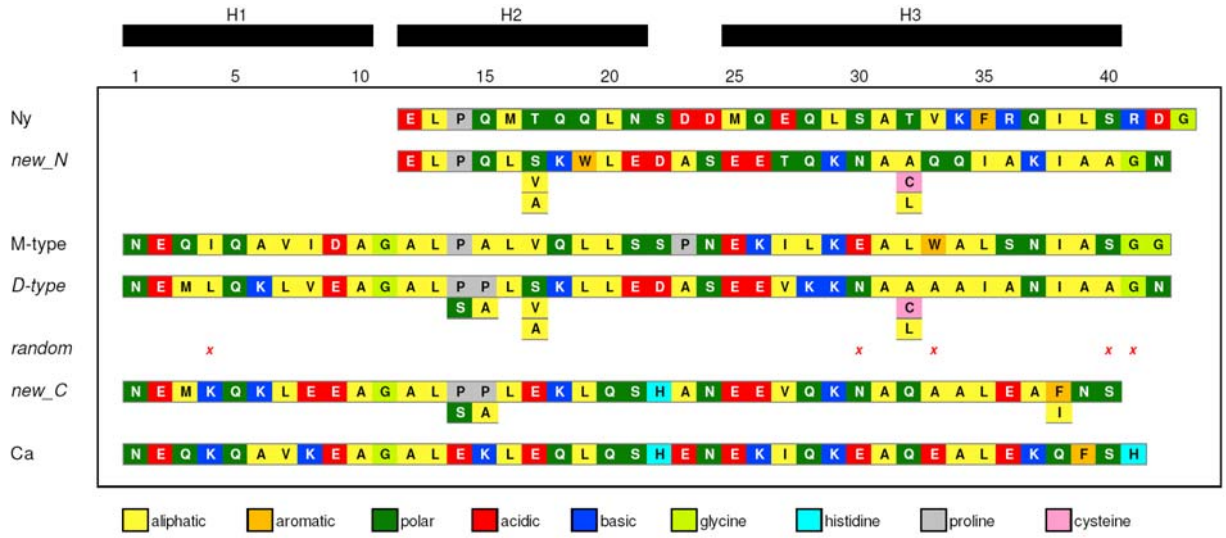


Fig. 3 New designed sequences and comparison with the former consensus design. D-type is the final internal module, after replacement of the serines at positions 33 and 36 by alanines. The new capping repeats, similarly, are aligned with the previously used. The additional residues below the designed sequences are putative mutations described in the text. The randomized positions of a library based on D-type module are indicated by the red x.

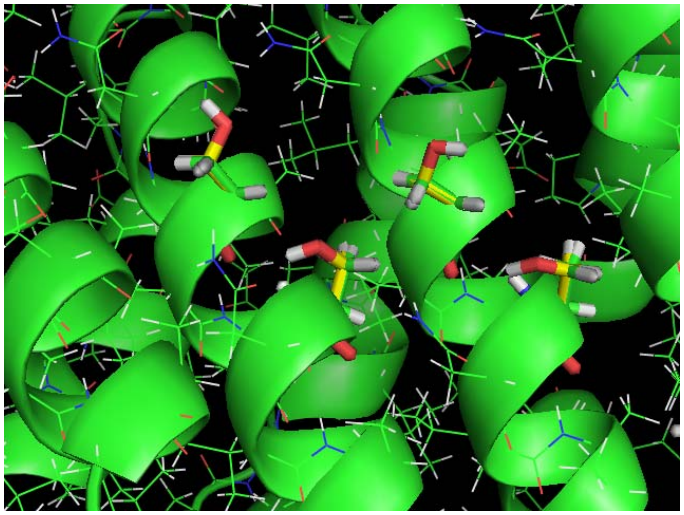


Fig. 4 Serines on H3 surface. Serines at positions 33 and 36 are depicted in yellow. They were replaced by alanine (in green, superimposed) following the results of molecular dynamics simulations that indicated their destabilizing effect.

Other residues suggested by ROSETTA can potentially be replaced. A double proline (positions 14-15, Fig. 5a) is a destabilizing and rather uncommon feature in an α-helix, even if sometimes observed in repeats belonging to natural armadillo proteins. Prolines do not possess backbone amine hydrogens to be involved in the hydrogen bond pattern characteristic of α-helices. The presence of one proline is tolerated at the beginning of a helix, where no hydrogen bond partner is available for a backbone amine hydrogen. Pro14 is highly conserved among the family but in certain cases a proline is placed at position 15, either in combination

with Pro14 or with other residues at that position. ROSETTA tends to introduce the Pro15 because the side chain fits in a pocket at the interface with the previous repeat. Pro14, in a solvent-exposed position, could be potentially replaced by serine, being at the very beginning of the helix and thus with less influence of its low helical propensity, whether Pro15 can only be substituted by an alanine to fit into the pocket. Long side chains were excluded as alternative to avoid additional effects (less likely with short side chains) which would mask the impact of proline replacement.

Ser 17 was proposed by ROSETTA because of its ability to form a conserved hydrogen bond with the backbone of Ala 10 in the following repeat (Fig. 5a). However, as mentioned above, serine does not possess a high helical propensity and in solution the hydrogen bond could be not present if the repeat orientation changes slightly or because of protein dynamics. Alanine is a valid alternative in terms of helical propensity, but also valine is a possible candidate because of its frequent occurrence at this position in armadillo proteins (Fig. 5b).

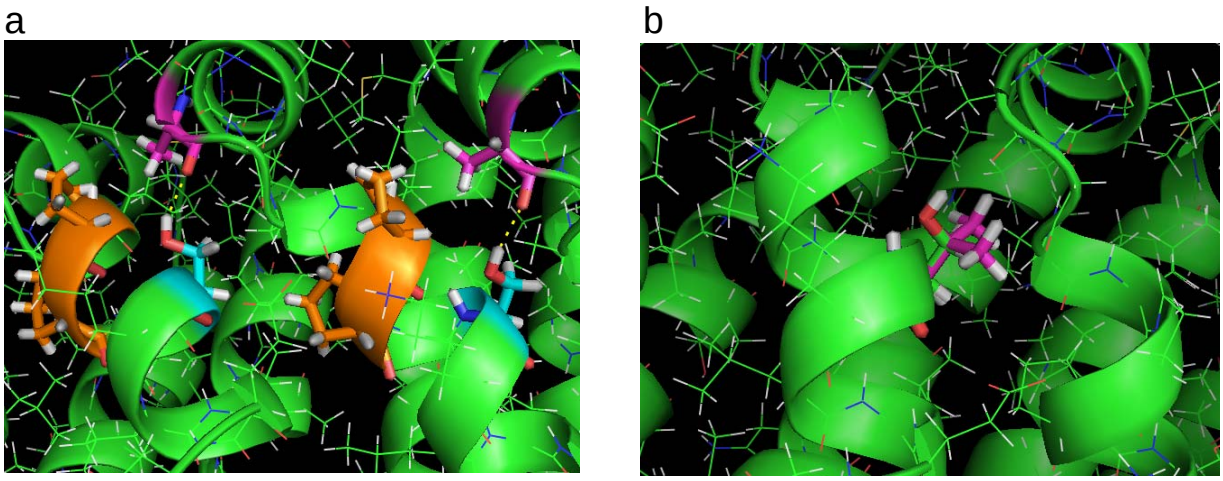


Fig. 5 Surface of helix 2. a Double proline at positions 14-15 (in orange) and serine at position 17 forming a hydrogen bond with backbone of ala 10 in the following repeat. b replacement of ser 17 with valine (in magenta).

Ala 32 is a residue pointing to a small cavity present at the interface between repeats (Fig. 6). This cavity could be filled, improving the packing, by a cysteine that was excluded during the sequence search to avoid possible problems in dimerization of the final protein. This amino acid is indeed used to fill exactly the same type of cavity in natural armadillo domain, and in particular in the repeats with geometry corresponding to our model. An alternative is provided by the highly conserve leucine, which could perfectly fill the cavity, but, by doing

so, would probably displace the neighboring helix 3, changing the repeat and the binding site geometry.

All the described mutations are allowed in the model structures with small adjustments. However, no additional information directing the choice toward a particular residue was provided by simulations; the mutants should be then expressed and compared with the original ROSETTA-based sequence. The mutations in the inner repeats will be present also on the N-cap (positions 17 and 32) or the C-cap (positions 14 and 15), because present at the interface between the capping repeat and the internal repeat.

Concerning the N-capping repeat, a tryptophan at position 19 (Fig. 7) has been introduced for practical reasons, to detect the protein by absorbance at 280 nm and easily calculate the concentration based on the extinction coefficient.

The phenylalanine at position 38 in the C-terminal capping repeat (Fig. 8) was introduced to seal the hydrophobic core as observed in the natural capping repeats. However, in natural proteins the phenylalanine is located at position 39. The geometry we have chosen for the designed molecule influences also the capping repeats. In our model, the position 38 is in a better conformation to provide the phenylalanine side chain than the position 39. Interestingly, without any additional information, ROSETTA was not able to propose a reasonable suggestion for the C-cap sequence, but after introducing a phenylalanine at position 38 all the rest was easily redesigned. From this result it appeared that also an isoleucine, present in this position in the internal repeats, could fit, representing a valid alternative to phenylalanine with small adjustments of backbone and side chains.

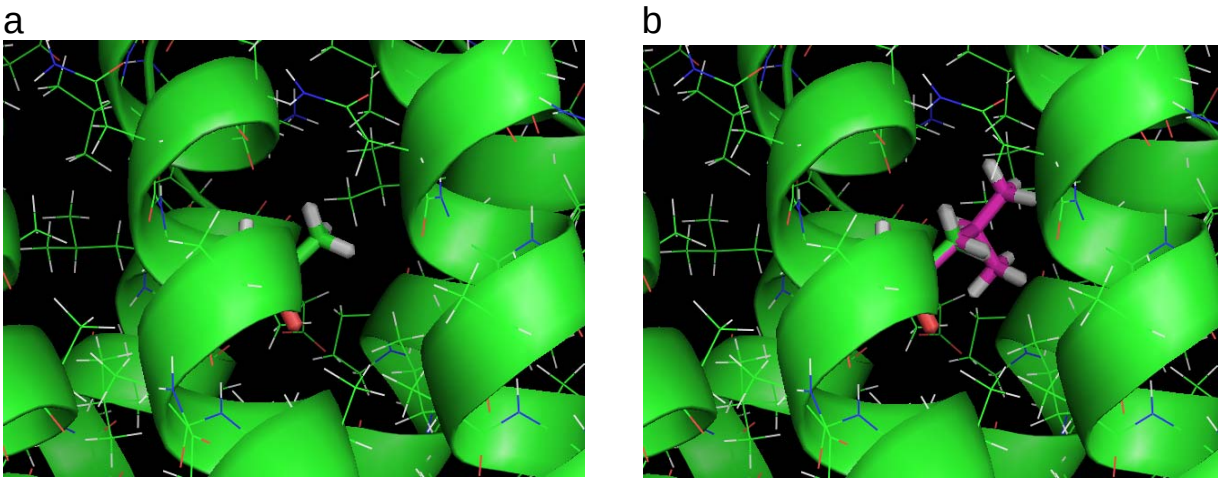


Fig. 6 Position 32 on helix 3. (a) Ala suggested by the ROSETTA design leaves a small cavity in the core. (b) The replacement by Leu, in magenta, fills the cavity, but push the following helix 3 away, most likely disrupting partially the geometry.

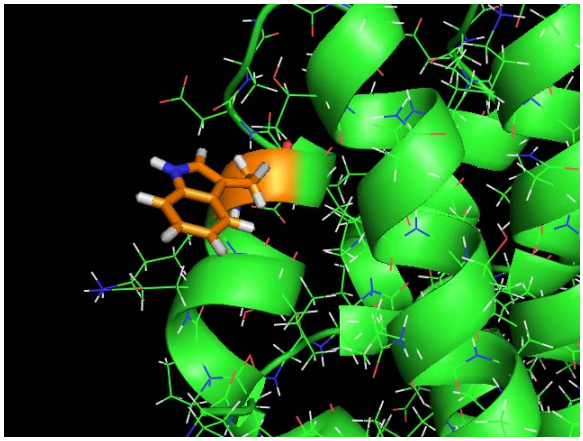


Fig. 7 Tryptophan in N-cap at position 19 is in an exposed position and ist property should not change upon unfolding of the protein.

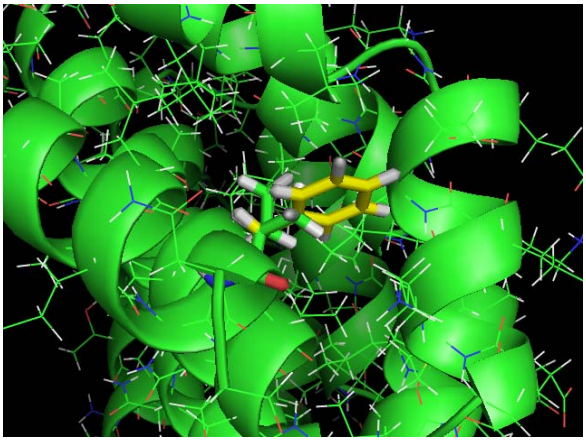


Fig. 8 Position 38 of C-cap. Phe (in yellow) and Ile (in green) are shown as possible alternatives in this position.

ROSETTA was used also to define the residues sterically allowed at the putative randomized positions (Fig. 3) and to avoid, in the generation of the next libraries, residues with the potential to disrupt significantly the structure. The evaluation of each randomized position was done by replacing the others randomized positions in the same repeat by alanines. Also all the randomized positions in the neighboring repeats were substituted by alanines. In the final list of allowed candidates Cys was excluded from all the positions to avoid formation of disulphide bonds and dimerization.

Out of the positions defined as to be randomized in Chapter 3, position 36 was dropped, because only Ser and Ala were sterically allowed, and Ser was already discarded after the results of molecular dynamic simulations. Position 4 was restricted to Ala, Asp, Glu, Asn, Lys, Arg, Ser; Gly is allowed from the calculation as well, but is probably going to introduce too much flexibility. The new scaffold does not allow β-branched residues as before at this position, like isoleucine and threonine. The main influence of position 4 seems to be more on the structure of the loop between H3 and H1, involved as well in binding, than on the direct recognition of the target.

As observed before, both positions 29 and 30 are used alternatively in natural armadillo proteins to bind side chains of the target peptide. However, position 30 is by far the most frequently used. In addition, most of the side chains occupying the position 29 would require a Ser in position 30, preventing it to be used as randomized position. Models of target peptides with long side chains are almost never reaching the residues at position 29 in the armadillo protein: only a few contacts could be possible if long side chains are present both in the target

and in the armadillo protein at this position. Therefore, in the working library position 29 will not be randomized. In contrast, position 30 can accommodate all residues, except Pro. Position 33, similarly, can accept all residues except Ile and Val, with intermediate indications for Thr. Position 40 accepts as well almost any residue, with the exception of Pro and Ile. Position 41 is restricted to Ala, Asp, Met, Asn, Gln, Ser, Thr and Val. In addition, replacement by Glu assumes a high flexibility of Met at position 3 and Lys can be positioned only with a strained rotamer and while keeping the charge quite buried. Asn at position 42 will be replaced in the library by Gly, to compensate the likely loss of flexibility associated with several of the side chains introduced with the randomization of position 41.

Toward experiments

All the variations in internal and capping repeats can, at this stage, be evaluated only at the experimental level by comparing the characteristics of proteins carrying different combination of mutations.

However, the effect of mutations at positions 14, 15, 17 is most likely independent on the changes at position 32. The two effects are then going to be additives in a first approximation. Taking this consideration into account, a sequential mutation strategy at the experimental level can be used to determine the optimal combination of residues in the first group (positions 14, 15 and 17 are spatially close and influence each other) and then to search for additional improvements by mutating the residue at position 32.

Since also the capping repeats contain these mutations, a first evaluation of internal repeats using Ny and Ca as capping repeats would reduce the number of combinations. The previously used capping repeats can, most likely, lead to soluble proteins as already shown for different types of modules [6] (chapter 2). The internal repeats will determine the protein characteristics and the best mutations will be introduced in the new capping repeats. Only one N-cap and two C-caps, containing phenylalanine or isoleucine, will be then tested.

The validity of the design is going to be evaluated soon and the best version will lead to a second generation library for the selection of peptide binders.

References

1. Rohl, C. A., Strauss, C. E., Misura, K. M. & Baker, D. (2004). Protein structure prediction using Rosetta. *Methods Enzymol* **383**, 66-93.
2. Simons, K. T., Kooperberg, C., Huang, E. & Baker, D. (1997). Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J Mol Biol* **268**, 209-25.
3. Kuhlman, B., Dantas, G., Ireton, G. C., Varani, G., Stoddard, B. L. & Baker, D. (2003). Design of a novel globular protein fold with atomic-level accuracy. *Science* **302**, 1364-8.
4. Rothlisberger, D., Khersonsky, O., Wollacott, A. M., Jiang, L., DeChancie, J., Betker, J., Gallaher, J. L., Althoff, E. A., Zanghellini, A., Dym, O., Albeck, S., Houk, K. N., Tawfik, D. S. & Baker, D. (2008). Kemp elimination catalysts by computational enzyme design. *Nature* **453**, 190-5.
5. Jiang, L., Althoff, E. A., Clemente, F. R., Doyle, L., Rothlisberger, D., Zanghellini, A., Gallaher, J. L., Betker, J. L., Tanaka, F., Barbas, C. F., 3rd, Hilvert, D., Houk, K. N., Stoddard, B. L. & Baker, D. (2008). De novo computational design of retro-aldol enzymes. *Science* **319**, 1387-91.
6. Parmeggiani, F., Pellarin, R., Larsen, A. P., Varadamsetty, G., Stumpp, M. T., Zerbe, O., Caflisch, A. & Pluckthun, A. (2008). Designed armadillo repeat proteins as general peptide-binding scaffolds: consensus design and computational optimization of the hydrophobic core. *J Mol Biol* **376**, 1282-304.

Conclusions

The way to peptide binders

The work described in this thesis was focused on the design of armadillo repeat proteins as potential peptide binders. The modularity and the conserved binding mode were the main reasons behind the choice of this protein family as starting framework. However, the generation of designed modules was the key point to achieve in order to exploit the modular arrangement. Natural armadillo proteins do not possess, indeed, conserved interfaces between the repeats that would favor the exchange, insertion and deletion of repeats without altering the overall structure.

Consensus design was an attractive strategy, taking advantage of the number of sequences available. And the devised rotamer sampling approach brought the designed proteins from a molten globule to a native-like state.

The main issues of a design project are the feasibility and the usefulness. It is not only important to design a protein but to provide properties, functions and applications not available or less attractive before. Armadillo repeat proteins have proven to be successfully designable and the preliminary results indicate that the selection of peptide binders is possible, even though more work still has to be done to turn these proteins in an efficient and reliable tool.

The impact of such designed proteins on the market will not be the same as for antibodies or, in more recent years, other alternative scaffolds, because of the narrower range of targets to

which they address. Most of the clinically relevant targets are folded proteins and other scaffold can more easily fit to this requirement.

It is in the diagnostic and research fields that armadillo repeat proteins could play a major role: screening, tissue profiling or detection of post-translational modification are all potential areas of use of designed armadillo repeat proteins, where the fast generation of specific binders and the low cost of production represent the key requirements.

Besides proof of principle selections, the current research is indeed directed towards commercially relevant areas where the specific recognition is required and where the modular structure of armadillo repeat proteins can represent an advantage, e.g. detection of post-translational modifications (phosphorylation, acetylation, methylation, etc …), discrimination of mutations in single residues, generation of dipeptide specific modules.

At the same time the construction of a second generation molecules is ongoing, based on the described computational design, to achieve higher control of the molecule shape and flexibility and to provide a geometrically ideal binding site for target peptides.

In parallel, structure determination is actively pursued, to clarify the three-dimensional structure of designed armadillo repeat proteins.

The work described here represents only the first step of a long-term project, the generation of specific binders without selection, but the results already achieved show that armadillo repeat proteins can be indeed designed, realized and refined and that they possess the characteristics and the potential to become peptide binders. This is not a trivial step considering the number of designed proteins that turn out as insoluble or molten globules. And a soluble, monomeric and stable protein is the requirement for the generation of a library of potential binders.

Appendices

Appendix 1

Determination of protein concentration

Protein concentration is usually determined with detection methods based on complex formation (Bradford, Biuret, Lowry or bicinchoninic acid (BCA) assays) or inherent spectral properties of proteins [1]. Among the second group, absorbance at 280 nm is the most widely used method for its simplicity, considering that extinction coefficient can be easily calculated for every protein. There is, however, a protein-to-protein variation, and this approach is not suitable for protein lacking aromatic groups and especially tryptophan that contribute for the majority of the absorption.

The concentration of designed armadillo repeat proteins can be easily determined when the internal modules possess at least one tryptophan, as for type-I, C, M. Type-T proteins do not contain any aromatic group and library members can be completely devoid of tryptophan, which occupy a randomizable position in the original type M framework. Phenylalanine is the only aromatic residue present in the capping repeat but does not contribute significantly to absorbance at 280 nm, making impossible to calculate an extinction coefficient.

Bradford and BCA assay were performing poorly in terms of reproducibility and comparison with absorption at 280 nm. The main reason is probably the use of a calibration curve based on an average reference protein, as BSA, but with remarkably different amino acid composition, which is known to contribute to the signal.

An alternative method was developed by Whitaker and Granum [2], with the aim of filtering out the effect of absorbing side chains and obtaining a concentration value dependent only on the protein backbone. The difference in absorption at 235 nm and 280 nm is in this case used to calculate the protein concentration according to equation (1).

$$\text{Concentration (mg/ml)} = (A_{235}-A_{280})/2.51 \quad (1)$$

The method is based on the fact that the absorbance at 280 nm of tryptophan, the strongest contributor to the extinction coefficient is equal, in value, to the absorbance at 235 nm. A

correction factor determined experimentally from a series of proteins is, however needed, to take into account different effects on absorbance present in the proteins.

This approach does not offer, generally, advantages in comparison to other methods but can be fully exploited in case of designed armadillo repeat proteins. The members containing tryptophan can be used to build a calibration curve allowing the calculation of a specific correction factor for the Whitaker-Granum method. Therefore, this approach becomes particularly relevant for designed armadillo repeat proteins missing tryptophan, because it allows a direct calculation of the concentration and it is compatible with the values determined by absorbance at 280 nm for the other proteins containing tryptophan.

Different samples and batches of C-type proteins containing 2, 4 or 8 internal repeats were used for the calculation of the correction factor (Fig. 1). Linear regression resulted in a correction factor (Cf) equal to 1.01 (to be multiplied for the difference A_{235}-A_{280}) and a correlation coefficient R^2=0.883. The modified version of the equation is:

$$\text{Concentration (mg/ml)} = (A_{235}-A_{280})\text{x}1.01 \quad (2)$$

If the two outliers are excluded Cf=0.95 and R^2=0.975.

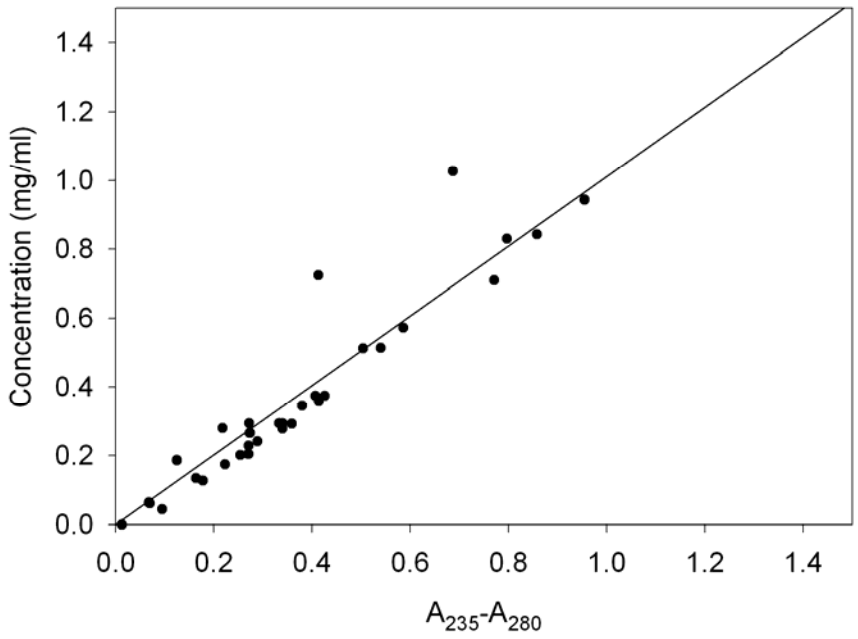


Fig.1 Calibration curve for the calculation of the correction factor. The dots represent experimental values of type C proteins. The concentrations were determined using the calculated extinction coefficient at 280 nm. The line indicates the linear regression curve, with intercept fixed to a value of 0.

The extinction coefficients were calculated using ProtParam [3], which is nowadays based on values for tryptophan and tyrosine determined by Pace *et al.* [4]. If extinction coefficients based on protected peptides in solution [5] are used, Cf=0.98 for the full set of data and Cf=0.92 when the outliers are excluded.

The equation (2) has been used in this work for the determination of the concentration of all the proteins without tryptophan. As for the original Whitaker-Granum method, substances strongly absorbing at 235 nm interfere with the measurements. Therefore, concentrations cannot be directly calculated for samples obtained after IMAC purification using Ni NTA material, due to the high imidazole concentration.

References

1. Simpson, R. J. (2004). Measuring the concentration of proteins. In *Purifying proteins for proteomics: a laboratory manual* (Simpson, R. J., ed.), pp. 659-700. Cold Spring Harbor Laboratory Press.
2. Whitaker, J. R. & Granum, P. E. (1980). An absolute method for protein determination based on difference in absorbance at 235 and 280 nm. *Anal Biochem* **109**, 156-9.
3. Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R. D. & Bairoch, A. (2003). ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res* **31**, 3784-8.
4. Pace, C. N., Vajdos, F., Fee, L., Grimsley, G. & Gray, T. (1995). How to measure and predict the molar absorption coefficient of a protein. *Protein Sci* **4**, 2411-23.
5. Edelhoch, H. (1967). Spectroscopic determination of tryptophan and tyrosine in proteins. *Biochemistry* **6**, 1948-54.

Appendix 2

List of armadillo crystal structures

PDB	name	organism	Res. (Å)	Year	ligand	Ref.
1BK5	importin α	S. cerevisiae	2.2	1998	/	1
1BK6	importin α	S. cerevisiae	2.8	1998	NLS sv40	1
1EE4	importin α	S. cerevisiae	2.1	2000	NLS c-myc	2
1EE5	importin α	S. cerevisiae	2.4	2000	NLS nucleoplasmin	2
1EJL	importin α	mouse	2.8	2000	NLS sv40	3
1EJY	importin α	mouse	2.9	2000	NLS nucleoplasmin	3
1G3J	β-catenin	human	2.1	2000	tcf3	4
1I7W	β-catenin	mouse	2.0	2001	e-cadherin P	5
1I7X	β-catenin	mouse	3.0	2001	e-cadherin	5
1IAL	importin α	mouse	2.5	1999	IBB domain linked	6
1IQ1	importin α	mouse	2.8	2001	IBB domain cut	7
1JDH	β-catenin	human	1.9	2001	tcf4	8
1JPP	β-catenin	mouse	3.1	2001	APC repeats 15	9
1JPW	β-catenin	human	2.5	2001	tcf4	10
1LUJ	β-catenin	human	2.5	2002	icat	11
1M1E	β-catenin	mouse	2.1	2002	icat	12
1PJM	importin α	mouse	2.5	2003	NLS rb	13
1PJN	importin α	mouse	2.5	2003	NLS nin2	13
1Q1S	importin α	mouse	2.3	2003	NLS sv40-P	14
1Q1T	importin α	mouse	2.5	2003	NLS sv40	14
1QZ7	β-catenin	human	2.2	2003	axin	15
1UN0	importin α	S. cerevisiae	2.6	2003	nup2p	16
2BCT	β-catenin	mouse	2.9	1997	/	17
3BCT	β-catenin	mouse	2.9	1997	/	17
1TH1	β-catenin	human	2.5	2004	APC repeats 20	18
1T08	β-catenin	human	2.1	2004	icat / APC rep	19
1V18	β-catenin	mouse	2.1	2004	APC 20mer phosp.	19
1WA5	importin α	S. cerevisiae	2	2004	ran / CSE1	20
1XM9	plakophilin	human	2.8	2005	/	21
1Y2A	importin α	mouse	2.2	2005	PLSCR1 NLS	22
2C1T	importin α	S. cerevisiae	2.6	2005	nup2	23
2JDQ	importin α	human	2.2	2007	PB2	24

References

1. Conti, E., Uy, M., Leighton, L., Blobel, G. & Kuriyan, J. (1998). Crystallographic analysis of the recognition of a nuclear localization signal by the nuclear import factor karyopherin alpha. *Cell* **94**, 193-204.
2. Conti, E. & Kuriyan, J. (2000). Crystallographic analysis of the specific yet versatile recognition of distinct nuclear localization signals by karyopherin alpha. *Structure Fold Des* **8**, 329-38.
3. Fontes, M. R., Teh, T. & Kobe, B. (2000). Structural basis of recognition of monopartite and bipartite nuclear localization sequences by mammalian importin-alpha. *J Mol Biol* **297**, 1183-94.
4. Graham, T. A., Weaver, C., Mao, F., Kimelman, D. & Xu, W. (2000). Crystal structure of a beta-catenin/Tcf complex. *Cell* **103**, 885-96.
5. Huber, A. H. & Weis, W. I. (2001). The structure of the beta-catenin/E-cadherin complex and the molecular basis of diverse ligand recognition by beta-catenin. *Cell* **105**, 391-402.
6. Kobe, B. (1999). Autoinhibition by an internal nuclear localization signal revealed by the crystal structure of mammalian importin alpha. *Nat Struct Biol* **6**, 388-97.
7. Catimel, B., Teh, T., Fontes, M. R., Jennings, I. G., Jans, D. A., Howlett, G. J., Nice, E. C. & Kobe, B. (2001). Biophysical characterization of interactions involving importin-alpha during nuclear import. *J Biol Chem* **276**, 34189-98.
8. Graham, T. A., Ferkey, D. M., Mao, F., Kimelman, D. & Xu, W. (2001). Tcf4 can specifically recognize beta-catenin using alternative conformations. *Nat Struct Biol* **8**, 1048-52.
9. Eklof Spink, K., Fridman, S. G. & Weis, W. I. (2001). Molecular mechanisms of beta-catenin recognition by adenomatous polyposis coli revealed by the structure of an APC-beta-catenin complex. *Embo J* **20**, 6203-12.
10. Poy, F., Lepourcelet, M., Shivdasani, R. A. & Eck, M. J. (2001). Structure of a human Tcf4-beta-catenin complex. *Nat Struct Biol* **8**, 1053-7.
11. Graham, T. A., Clements, W. K., Kimelman, D. & Xu, W. (2002). The crystal structure of the beta-catenin/ICAT complex reveals the inhibitory mechanism of ICAT. *Mol Cell* **10**, 563-71.
12. Daniels, D. L. & Weis, W. I. (2002). ICAT inhibits beta-catenin binding to Tcf/Lef-family transcription factors and the general coactivator p300 using independent structural modules. *Mol Cell* **10**, 573-84.
13. Fontes, M. R., Teh, T., Jans, D., Brinkworth, R. I. & Kobe, B. (2003). Structural basis for the specificity of bipartite nuclear localization sequence binding by importin-alpha. *J Biol Chem* **278**, 27981-7.
14. Fontes, M. R., Teh, T., Toth, G., John, A., Pavo, I., Jans, D. A. & Kobe, B. (2003). Role of flanking sequences and phosphorylation in the recognition of the simian-virus-40 large T-antigen nuclear localization sequences by importin-alpha. *Biochem J* **375**, 339-49.
15. Xing, Y., Clements, W. K., Kimelman, D. & Xu, W. (2003). Crystal structure of a beta-catenin/axin complex suggests a mechanism for the beta-catenin destruction complex. *Genes Dev* **17**, 2753-64.
16. Matsuura, Y., Lange, A., Harreman, M. T., Corbett, A. H. & Stewart, M. (2003). Structural basis for Nup2p function in cargo release and karyopherin recycling in nuclear import. *Embo J* **22**, 5358-69.
17. Huber, A. H., Nelson, W. J. & Weis, W. I. (1997). Three-dimensional structure of the armadillo repeat region of beta-catenin. *Cell* **90**, 871-82.
18. Xing, Y., Clements, W. K., Le Trong, I., Hinds, T. R., Stenkamp, R., Kimelman, D. & Xu, W. (2004). Crystal structure of a beta-catenin/APC complex reveals a critical role for APC phosphorylation in APC function. *Mol Cell* **15**, 523-33.
19. Ha, N. C., Tonozuka, T., Stamos, J. L., Choi, H. J. & Weis, W. I. (2004). Mechanism of phosphorylation-dependent binding of APC to beta-catenin and its role in beta-catenin degradation. *Mol Cell* **15**, 511-21.
20. Matsuura, Y. & Stewart, M. (2004). Structural basis for the assembly of a nuclear export complex. *Nature* **432**, 872-7.
21. Choi, H. J. & Weis, W. I. (2005). Structure of the armadillo repeat domain of plakophilin 1. *J Mol Biol* **346**, 367-76.
22. Chen, M. H., Ben-Efraim, I., Mitrousis, G., Walker-Kopp, N., Sims, P. J. & Cingolani, G. (2005). Phospholipid scramblase 1 contains a nonclassical nuclear localization signal with unique binding site in importin α. *J Biol Chem* **280**, 10599-606.
23. Matsuura, Y. & Stewart, M. (2005). Nup50/Npap60 function in nuclear protein import complex disassembly and importin recycling. *Embo J* **24**, 3681-9.
24. Tarendeau, F., Boudet, J., Guilligay, D., Mas, P. J., Bougault, C. M., Boulo, S., Baudin, F., Ruigrok, R. W., Daigle, N., Ellenberg, J., Cusack, S., Simorre, J. P. & Hart, D. J. (2007). Structure and nuclear import function of the C-terminal domain of influenza virus polymerase PB2 subunit. *Nat Struct Mol Biol* **14**, 229-33.

Appendix 3

Oligonucleotides

The oligonucleotides listed here were used for the generation of the generation of KQ, QK and QQ molecules, for the synthesis of the libraries, for ribosome display and for improvement of the pRDV vector. The oligonucleotides used for cloning of natural armadillo repeat proteins, and for generation of the consensus repeats are listed in the supplementary materials of Parmeggiani et al. [1] and reported in Chapter 2, page 77.

Oligonucleotides

name	sequence 5'-3' direction	description (for=forward, rev=reverse)
KQfor	CCGAACGAGAAGATCCTGCAAGAAGCTC TGTGGGC	for mutation KK->KQ
KQrev	GCCCACAGAGCTTCTTGCAGGATCTTCT CGTTCGG	rev mutation KK->KQ
QKfor	CCTCTCCGAACGAGCAGATCCTGAAAGA AGC	for mutation KK->QK
QKrev	GCTTCTTTCAGGATCTGCTCGTTCGGAG AGG	rev mutation KK->QK
QQfor	CCTCTCCGAACGAGCAGATCCTGCAAGA AGCTCTGTGGGC	for mutation KK->QQ
QQrev	GCCCACAGAGCTTCTTGCAGGATCTGCT CGTTCGGAGAGG	rev mutation KK->QQ
Ny4libRbis	TTCCTGGTACCCTAAGGTCTCATTCGTT ACCATCACGAGACAGGATCTG	rev replacement for assembly Ny cap in library format
Ca1libF	CCAGGGATCCTCTAGATAGGAAGACCTC GAACAGAAACAGGC	for replacement for assembly Ca cap in library format
Ca2libR	GTTTCTCCAGAGCACCAGCTTCTTTAAC AGCCTGTTTCTGTTCGAGGTC	rev replacement for assembly Ca cap in library format
lib1F1	CCAGGGATCCTAGGAAGACCTCGAACAA AYCCAAGCTGTTATCG	for assembly libraries, randomized position 4
lib1F2	CCAGGGATCCTAGGAAGACCTCGAACAA VAACAAGCTGTTATCG	for assembly libraries, randomized position 4
lib1F3	CCAGGGATCCTAGGAAGACCTCGAACAA CRCCAAGCTGTTATCG	for assembly libraries, randomized position 4
lib2R	CCAGAGCCGGCAGAGCACCAGCATCGAT AACAGCTTG	rev assembly libraries
lib3F	CTGCCGGCTCTGGTTCAACTGCTGTCCT CTC	for assembly libraries
lib4R	TTTCAGGATCTTCTCGTTCGGAGAGGAC AGCAG	rev assembly KK library
lib5F	CGAGAAGATCCTGAAANNNGCTCTGNNN GCTCTGNNNAACATCGCTNNNNNNGGTA ACGAATGAG	for assembly KK library, randomized positions 30,33,36,40,41 with trinucleotides

lib6R	TTCCTGGTACCCTAAGGTCTCATTCGTTACC	rev assembly libraries
libFOR	CCAGGGATCCTAGGAAGACCTCGAAC	for general for library amplification
5KQ	CGAGAAGATCCTGCAA	for conversion KK library to KQ library
5QK	CGAGCAGATCCTGAAA	for conversion KK library to QK library
5QQ	CGAGCAGATCCTGCAA	for conversion KK library to QQ library
lib4RKQ	TTGCAGGATCTTCTCGTTCGGAGAGGACAGCAG	rev conversion KK library to KQ library
lib4RQK	TTTCAGGATCTGCTCGTTCGGAGAGGACAGCAG	rev conversion KK library to QK library
lib4RQQ	TTGCAGGATCTGCTCGTTCGGAGAGGACAGCAG	rev conversion KK library to QQ library
Ca6RhindRD	TTCCTAAGCTTGTGGGAGAACTGCTTCTCCAGAGCTTCC	rev amplification to remove stop codon for ribosome display
T7b	ATACGAAATTAATACGACTCACTATAGGGAGACCACAACGG	for outer primer ribosome display
tolAk	CCGCACACCAGTAAGGTGTGCGGTTTCAGTTGCCGCTTTCTTTCT	rev outer primer ribosome display
pRDVhis-for	GCATCACCATCACCATCACGG	for inner primer ribosome display
pRDVhis-rev	CCCCGAGGCCATATAAAGC	rev inner primer ribosome display
MRGfor	GGAGTCATGAGAGGATCGCATCACCATCACC	for modification pRDV into pRDVhis
MRGrev	CACAGGATCCGTGATGGTGATGGTGATGCG	rev modification pRDV into pRDVhis
3mCAGfor	GGAGATATATTCATGCGAGGGTCGCATCACCATC	for silent mutation CAG in pRDVhis for translation
3mCAGrev	GATGGTGATGCGACCCTCGCATGAATATATCTCC	rev silent mutation CAG in pRDVhis for translation
3mCCCfor	GGAGATATATTCATGCGCGGCTCGCATCACCATC	for silent mutation CCC in pRDVhis for translation
3mCCCrev	GATGGTGATGCGAGCCGCGCATGAATATATCTCC	rev silent mutation CCC in pRDVhis for translation

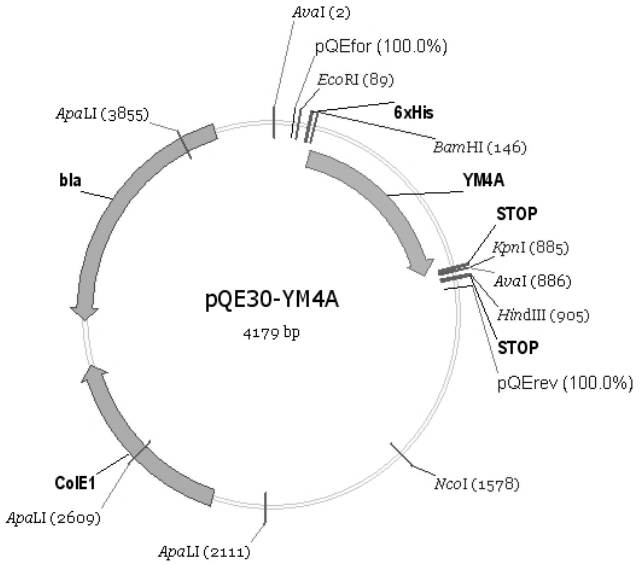
References

1. Parmeggiani, F., Pellarin, R., Larsen, A. P., Varadamsetty, G., Stumpp, M. T., Zerbe, O., Caflisch, A. & Pluckthun, A. (2008). Designed armadillo repeat proteins as general peptide-binding scaffolds: consensus design and computational optimization of the hydrophobic core. *J Mol Biol* **376**, 1282-304.

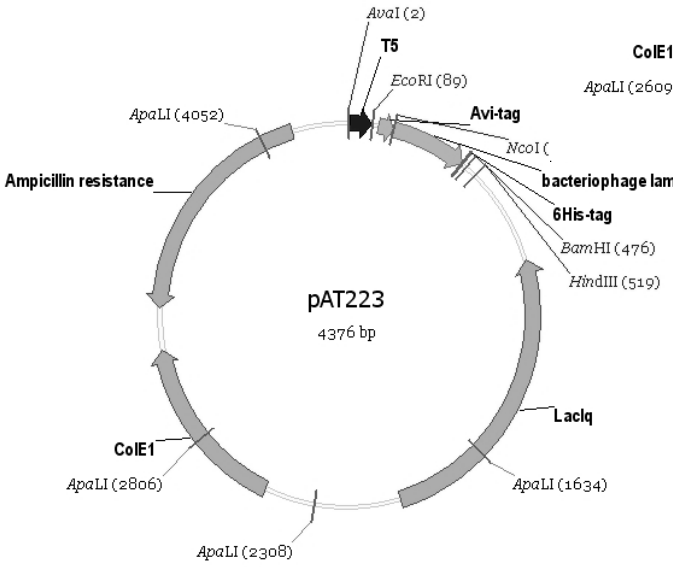
Appendix 4

Plasmids

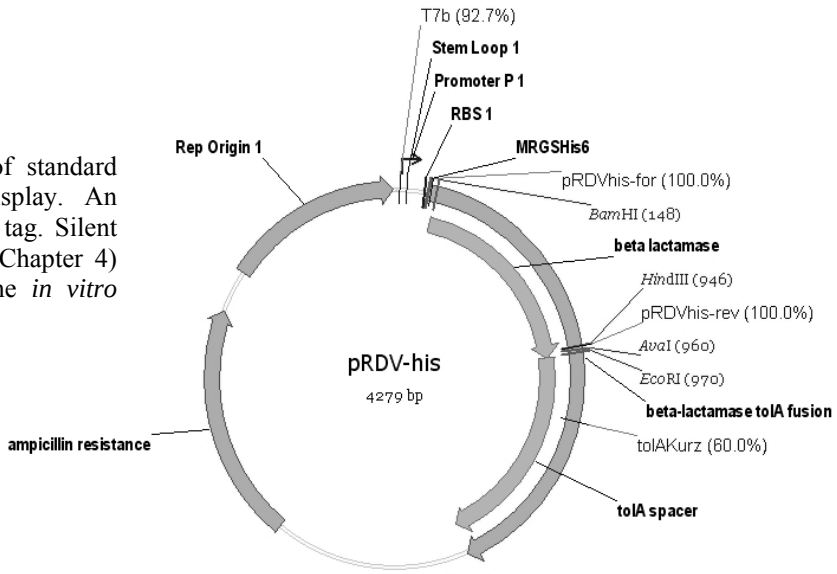
pQE30-YM4A: vector used in this work for the expression of all the armadillo repeat proteins. Contains an $MRGSH_6$ tag for purification at N-terminus of the protein. YM_4A is inserted in the vector in this case.



pAT223: vector used for the expression of fusion proteins, used in ELISA. Inserted sequences are fused to protein D (provided by the vector). for ELISAall the armadillo repeat proteins. Contains a H_6 tag for purification and an avitag for biotinylation.



pRDV-his: modified version of standard pRDVvector for ribosome display. An $MRGSH_6$ tag replaces the flag tag. Silent mutations (CAG, described in Chapter 4) were introduced to increase the *in vitro* translation efficiency.



Appendix 5

Designed proteins

Consensus designed N2Cs

The proteins contain 169 residues.
For C and I proteins the extinction coefficient at 280 nm (ε) is 11000 M^{-1} cm^{-1}.
T proteins do not possess any Trp, Tyr or Cys and ε cannot be calculated.

YI_2Y
MRGSHHHHHHGSELPQMTQQLNSDDMQEQLSATVKFRQILSRDG
NEQIQAVIDAGALPVLVELLSSPDNKIQKEALWALSNITSGG
NEQIQAVIDAGALPVLVELLSSPDNKIQKEALWALSNITSGG
DNINENADFIEKAGGMEKIFNCQQNENDKIYEKAYKIIETYFG
pI/Mw: 4.85 / 18910.14

YC_2Y
MRGSHHHHHHGSELPQMTQQLNSDDMQEQLSATVKFRQILSRDG
NEQIQAVIDAGGLPALVQLLSSPNEKILKEAAWALSNLASGG
NEQIQAVIDAGGLPALVQLLSSPNEKILKEAAWALSNLASGG
DNINENADFIEKAGGMEKIFNCQQNENDKIYEKAYKIIETYFG
pI/Mw: 5.04 / 18677.91

YI_2A
MRGSHHHHHHGSELPQMTQQLNSDDMQEQLSATVKFRQILSRDG
NEQIQAVIDAGALPVLVELLSSPDNKIQKEALWALSNITSGG
NEQIQAVIDAGALPVLVELLSSPDNKIQKEALWALSNITSGG
NEQKQAVKEAGALEKLEQLQSHENEKIQKEAQEALEKQFSH
pI/Mw: 5.19 / 18607.78

YC_2A
MRGSHHHHHHGSELPQMTQQLNSDDMQEQLSATVKFRQILSRDG
NEQIQAVIDAGGLPALVQLLSSPNEKILKEAAWALSNLASGG
NEQIQAVIDAGGLPALVQLLSSPNEKILKEAAWALSNLASGG
NEQKQAVKEAGALEKLEQLQSHENEKIQKEAQEALEKQFSH
pI/Mw: 5.44 / 18375.55

AI_2Y
MRGSHHHHHHGSSLNELVKQLNSDDQKQLKEAAQKLRQLASDG
NEQIQAVIDAGALPVLVELLSSPDNKIQKEALWALSNITSGG
NEQIQAVIDAGALPVLVELLSSPDNKIQKEALWALSNITSGG
DNINENADFIEKAGGMEKIFNCQQNENDKIYEKAYKIIETYFG
pI/Mw: 5.06 / 18670.87

AC_2Y
MRGSHHHHHHGSSLNELVKQLNSDDQKQLKEAAQKLRQLASDG
NEQIQAVIDAGGLPALVQLLSSPNEKILKEAAWALSNLASGG
NEQIQAVIDAGGLPALVQLLSSPNEKILKEAAWALSNLASGG
DNINENADFIEKAGGMEKIFNCQQNENDKIYEKAYKIIETYFG
pI/Mw: 5.32 / 18438.64

AI$_2$A
MRGSHHHHHHGSSLNELVKQLNSDDQKQLKEAAQKLRQLASDG
NEQIQAVIDAGALPVLVELLSSPDNKIQKEALWALSNITSGG
NEQIQAVIDAGALPVLVELLSSPDNKIQKEALWALSNITSGG
NEQKQAVKEAGALEKLEQLQSHENEKIQKEAQEALEKQFSH
pI/Mw: 5.45 / 18368.52

AC$_2$A
MRGSHHHHHHGSSLNELVKQLNSDDQKQLKEAAQKLRQLASDG
NEQIQAVIDAGGLPALVQLLSSPNEKILKEAAWALSNLASGG
NEQIQAVIDAGGLPALVQLLSSPNEKILKEAAWALSNLASGG
NEQKQAVKEAGALEKLEQLQSHENEKIQKEAQEALEKQFSH
pI/Mw: 5.73 / 18136.28

AC$_2$A modified Ser->Tyr
MRGSHHHHHHGSYLNELVKQLNSDDQKQLKEAAQKLRQLASDG
NEQIQAVIDAGGLPALVQLLSSPNEKILKEAAWALSNLASGG
NEQIQAVIDAGGLPALVQLLSSPNEKILKEAAWALSNLASGG
NEQKQAVKEAGALEKLEQLQSHENEKIQKEAQEALEKQFSH
pI/Mw: 5.73 / 18212.3 ε: 12490

YT$_2$A
MRGSHHHHHHGSELPQMTQQLNSDDMQEQLSATVKFRQILSRDG
EANKLAIRESGGIPALVRLLSSNNEKILEAATGTLHNLALHG
EANKLAIRESGGIPALVRLLSSNNEKILEAATGTLHNLALHG
NEQKQAVKEAGALEKLEQLQSHENEKIQKEAQEALEKQFSH
pI/Mw: 6.30 / 18585.8

YT$_2$Y
MRGSHHHHHHGSELPQMTQQLNSDDMQEQLSATVKFRQILSRDG
EANKLAIRESGGIPALVRLLSSNNEKILEAATGTLHNLALHG
EANKLAIRESGGIPALVRLLSSNNEKILEAATGTLHNLALHG
DNINENADFIEKAGGMEKIFNCQQNENDKIYEKAYKIIETYFG
pI/Mw: 6.04 / 18888.20

AT$_2$A
MRGSHHHHHHGSSLNELVKQLNSDDQKQLKEAAQKLRQLASDG
EANKLAIRESGGIPALVRLLSSNNEKILEAATGTLHNLALHG
EANKLAIRESGGIPALVRLLSSNNEKILEAATGTLHNLALHG
NEQKQAVKEAGALEKLEQLQSHENEKIQKEAQEALEKQFSH
pI/Mw: 6.63 / 18346.57

AT$_2$Y
MRGSHHHHHHGSSLNELVKQLNSDDQKQLKEAAQKLRQLASDG
EANKLAIRESGGIPALVRLLSSNNEKILEAATGTLHNLALHG
EANKLAIRESGGIPALVRLLSSNNEKILEAATGTLHNLALHG
DNINENADFIEKAGGMEKIFNCQQNENDKIYEKAYKIIETYFG
pI/Mw: 6.36 / 18648.93

Consensus designed N4Cs

The proteins contain 253 residues.

For C and I proteins the extinction coefficient at 280 nm (ε) is 22000 M^{-1} cm^{-1}.

T proteins do not possess any Trp, Tyr or Cys and ε cannot be calculated.

YI_4Y

MRGSHHHHHHGSELPQMTQQLNSDDMQEQLSATVKFRQILSRDG
NEQIQAVIDAGALPVLVELLSSPDNKIQKEALWALSNITSGG
NEQIQAVIDAGALPVLVELLSSPDNKIQKEALWALSNITSGG
NEQIQAVIDAGALPVLVELLSSPDNKIQKEALWALSNITSGG
NEQIQAVIDAGALPVLVELLSSPDNKIQKEALWALSNITSGG
DNINENADFIEKAGGMEKIFNCQQNENDKIYEKAYKIIETYFG

pI/Mw: 4.61 / 27682.10

YC_4Y

MRGSHHHHHHGSELPQMTQQLNSDDMQEQLSATVKFRQILSRDG
NEQIQAVIDAGGLPALVQLLSSPNEKILKEAAWALSNLASGG
NEQIQAVIDAGGLPALVQLLSSPNEKILKEAAWALSNLASGG
NEQIQAVIDAGGLPALVQLLSSPNEKILKEAAWALSNLASGG
NEQIQAVIDAGGLPALVQLLSSPNEKILKEAAWALSNLASGG
DNINENADFIEKAGGMEKIFNCQQNENDKIYEKAYKIIETYFG

pI/Mw: 4.83 / 27217.63

YI_4A

MRGSHHHHHHGSELPQMTQQLNSDDMQEQLSATVKFRQILSRDG
NEQIQAVIDAGALPVLVELLSSPDNKIQKEALWALSNITSGG
NEQIQAVIDAGALPVLVELLSSPDNKIQKEALWALSNITSGG
NEQIQAVIDAGALPVLVELLSSPDNKIQKEALWALSNITSGG
NEQIQAVIDAGALPVLVELLSSPDNKIQKEALWALSNITSGG
NEQKQAVKEAGALEKLEQLQSHENEKIQKEAQEALEKQFSH

pI/Mw: 4.82 / 27379.74

YC_4A

MRGSHHHHHHGSELPQMTQQLNSDDMQEQLSATVKFRQILSRDG
NEQIQAVIDAGGLPALVQLLSSPNEKILKEAAWALSNLASGG
NEQIQAVIDAGGLPALVQLLSSPNEKILKEAAWALSNLASGG
NEQIQAVIDAGGLPALVQLLSSPNEKILKEAAWALSNLASGG
NEQIQAVIDAGGLPALVQLLSSPNEKILKEAAWALSNLASGG
NEQKQAVKEAGALEKLEQLQSHENEKIQKEAQEALEKQFSH

pI/Mw: 5.09 / 26915.28

AI_4Y

MRGSHHHHHHGSSLNELVKQLNSDDQKQLKEAAQKLRQLASDG
NEQIQAVIDAGALPVLVELLSSPDNKIQKEALWALSNITSGG
NEQIQAVIDAGALPVLVELLSSPDNKIQKEALWALSNITSGG
NEQIQAVIDAGALPVLVELLSSPDNKIQKEALWALSNITSGG
NEQIQAVIDAGALPVLVELLSSPDNKIQKEALWALSNITSGG
DNINENADFIEKAGGMEKIFNCQQNENDKIYEKAYKIIETYFG

pI/Mw: 4.73 / 27442.84

AC_4Y

MRGSHHHHHHGSSLNELVKQLNSDDQKQLKEAAQKLRQLASDG
NEQIQAVIDAGGLPALVQLLSSPNEKILKEAAWALSNLASGG
NEQIQAVIDAGGLPALVQLLSSPNEKILKEAAWALSNLASGG
NEQIQAVIDAGGLPALVQLLSSPNEKILKEAAWALSNLASGG
NEQIQAVIDAGGLPALVQLLSSPNEKILKEAAWALSNLASGG
DNINENADFIEKAGGMEKIFNCQQNENDKIYEKAYKIIETYFG

pI/Mw: 4.99 / 26978.37

AI_4A
MRGSHHHHHHGSSLNELVKQLNSDDQKQLKEAAQKLRQLASDG
NEQIQAVIDAGALPVLVELLSSPDNKIQKEALWALSNITSGG
NEQIQAVIDAGALPVLVELLSSPDNKIQKEALWALSNITSGG
NEQIQAVIDAGALPVLVELLSSPDNKIQKEALWALSNITSGG
NEQIQAVIDAGALPVLVELLSSPDNKIQKEALWALSNITSGG
NEQKQAVKEAGALEKLEQLQSHENEKIQKEAQEALEKQFSH
pI/Mw: 4.96 / 27140.48

AC_4A
MRGSHHHHHHGSSLNELVKQLNSDDQKQLKEAAQKLRQLASDG
NEQIQAVIDAGGLPALVQLLSSPNEKILKEAAWALSNLASGG
NEQIQAVIDAGGLPALVQLLSSPNEKILKEAAWALSNLASGG
NEQIQAVIDAGGLPALVQLLSSPNEKILKEAAWALSNLASGG
NEQIQAVIDAGGLPALVQLLSSPNEKILKEAAWALSNLASGG
NEQKQAVKEAGALEKLEQLQSHENEKIQKEAQEALEKQFSH
pI/Mw: 5.28 / 26676.01

YT_4A
MRGSHHHHHHGSELPQMTQQLNSDDMQEQLSATVKFRQILSRDG
EANKLAIRESGGIPALVRLLSSNNEKILEAATGTLHNLALHG
EANKLAIRESGGIPALVRLLSSNNEKILEAATGTLHNLALHG
EANKLAIRESGGIPALVRLLSSNNEKILEAATGTLHNLALHG
EANKLAIRESGGIPALVRLLSSNNEKILEAATGTLHNLALHG
NEQKQAVKEAGALEKLEQLQSHENEKIQKEAQEALEKQFSH
pI/Mw: 6.46 / 27335.8 no extinction coeff. at 280 (no trp, tyr, cys)

YT_4Y
MRGSHHHHHHGSELPQMTQQLNSDDMQEQLSATVKFRQILSRDG
EANKLAIRESGGIPALVRLLSSNNEKILEAATGTLHNLALHG
EANKLAIRESGGIPALVRLLSSNNEKILEAATGTLHNLALHG
EANKLAIRESGGIPALVRLLSSNNEKILEAATGTLHNLALHG
EANKLAIRESGGIPALVRLLSSNNEKILEAATGTLHNLALHG
DNINENADFIEKAGGMEKIFNCQQNENDKIYEKAYKIIETYFG
pI/Mw: 6.27 / 27638.22

AT_4A
MRGSHHHHHHGSSLNELVKQLNSDDQKQLKEAAQKLRQLASDG
EANKLAIRESGGIPALVRLLSSNNEKILEAATGTLHNLALHG
EANKLAIRESGGIPALVRLLSSNNEKILEAATGTLHNLALHG
EANKLAIRESGGIPALVRLLSSNNEKILEAATGTLHNLALHG
EANKLAIRESGGIPALVRLLSSNNEKILEAATGTLHNLALHG
NEQKQAVKEAGALEKLEQLQSHENEKIQKEAQEALEKQFSH
pI/Mw: 6.77 / 27096.60

AT_4Y
MRGSHHHHHHGSSLNELVKQLNSDDQKQLKEAAQKLRQLASDG
EANKLAIRESGGIPALVRLLSSNNEKILEAATGTLHNLALHG
EANKLAIRESGGIPALVRLLSSNNEKILEAATGTLHNLALHG
EANKLAIRESGGIPALVRLLSSNNEKILEAATGTLHNLALHG
EANKLAIRESGGIPALVRLLSSNNEKILEAATGTLHNLALHG
DNINENADFIEKAGGMEKIFNCQQNENDKIYEKAYKIIETYFG
pI/Mw: 6.54 / 27398.95

Consensus designed N8Cs

The proteins contain 421 residues.

For C and I proteins the extinction coefficient at 280 nm (ε) is 44000 M^{-1} cm^{-1}.

T proteins do not possess any Trp, Tyr or Cys and ε cannot be calculated.

YI_8A

MRGSHHHHHHGSELPQMTQQLNSDDMQEQLSATVKFRQILSRDG
NEQIQAVIDAGALPVLVELLSSPDNKIQKEALWALSNITSGG
NEQIQAVIDAGALPVLVELLSSPDNKIQKEALWALSNITSGG
NEQIQAVIDAGALPVLVELLSSPDNKIQKEALWALSNITSGG
NEQIQAVIDAGALPVLVELLSSPDNKIQKEALWALSNITSGG
NEQIQAVIDAGALPVLVELLSSPDNKIQKEALWALSNITSGG
NEQIQAVIDAGALPVLVELLSSPDNKIQKEALWALSNITSGG
NEQIQAVIDAGALPVLVELLSSPDNKIQKEALWALSNITSGG
NEQIQAVIDAGALPVLVELLSSPDNKIQKEALWALSNITSGG
NEQKQAVKEAGALEKLEQLQSHENEKIQKEAQEALEKQFSH

pI/Mw: 4.55 / 44923.67

YC_8A

MRGSHHHHHHGSELPQMTQQLNSDDMQEQLSATVKFRQILSRDG
NEQIQAVIDAGGLPALVQLLSSPNEKILKEAAWALSNLASGG
NEQIQAVIDAGGLPALVQLLSSPNEKILKEAAWALSNLASGG
NEQIQAVIDAGGLPALVQLLSSPNEKILKEAAWALSNLASGG
NEQIQAVIDAGGLPALVQLLSSPNEKILKEAAWALSNLASGG
NEQIQAVIDAGGLPALVQLLSSPNEKILKEAAWALSNLASGG
NEQIQAVIDAGGLPALVQLLSSPNEKILKEAAWALSNLASGG
NEQIQAVIDAGGLPALVQLLSSPNEKILKEAAWALSNLASGG
NEQIQAVIDAGGLPALVQLLSSPNEKILKEAAWALSNLASGG
NEQKQAVKEAGALEKLEQLQSHENEKIQKEAQEALEKQFSH

pI/Mw: 4.80 / 43994.73

YT_8A

MRGSHHHHHHGSELPQMTQQLNSDDMQEQLSATVKFRQILSRDG
EANKLAIRESGGIPALVRLLSSNNEKILEAATGTLHNLALHG
EANKLAIRESGGIPALVRLLSSNNEKILEAATGTLHNLALHG
EANKLAIRESGGIPALVRLLSSNNEKILEAATGTLHNLALHG
EANKLAIRESGGIPALVRLLSSNNEKILEAATGTLHNLALHG
EANKLAIRESGGIPALVRLLSSNNEKILEAATGTLHNLALHG
EANKLAIRESGGIPALVRLLSSNNEKILEAATGTLHNLALHG
EANKLAIRESGGIPALVRLLSSNNEKILEAATGTLHNLALHG
EANKLAIRESGGIPALVRLLSSNNEKILEAATGTLHNLALHG
NEQKQAVKEAGALEKLEQLQSHENEKIQKEAQEALEKQFSH

pI/Mw: 6.67 / 44835.90

YC_4A core mutants

The proteins contain 253 residues.

The extinction coefficient at 280 nm (ε) is 22000 M^{-1} cm^{-1} for all the proteins.

Mutant 1

MRGSHHHHHHGSELPQMTQQLNSDDMQEQLSATVKFRQILSRDG
NEQIQAVIDAGGLPALVQLLSSPNEKLLKEALWALSNLASGG
NEQIQAVIDAGGLPALVQLLSSPNEKLLKEALWALSNLASGG
NEQIQAVIDAGGLPALVQLLSSPNEKLLKEALWALSNLASGG
NEQIQAVIDAGGLPALVQLLSSPNEKLLKEALWALSNLASGG
NEQKQAVKEAGALEKLEQLQSHENEKIQKEAQEALEKQFSH
pI/MW: 5.09 / 27083.6

Mutant 2

MRGSHHHHHHGSELPQMTQQLNSDDMQEQLSATVKFRQILSRDG
NEQIQAVIDAGALPALVQLLSSPNEKVLKEALWALSNLASGG
NEQIQAVIDAGALPALVQLLSSPNEKVLKEALWALSNLASGG
NEQIQAVIDAGALPALVQLLSSPNEKVLKEALWALSNLASGG
NEQIQAVIDAGALPALVQLLSSPNEKVLKEALWALSNLASGG
NEQKQAVKEAGALEKLEQLQSHENEKIQKEAQEALEKQFSH
pI/MW: 5.09 / 27083.6

Mutant 3

MRGSHHHHHHGSELPQMTQQLNSDDMQEQLSATVKFRQILSRDG
NEQIQAVIDAGALPALVQLLSSPNEKLLKEALWALSNLASGG
NEQIQAVIDAGALPALVQLLSSPNEKLLKEALWALSNLASGG
NEQIQAVIDAGALPALVQLLSSPNEKLLKEALWALSNLASGG
NEQIQAVIDAGALPALVQLLSSPNEKLLKEALWALSNLASGG
NEQKQAVKEAGALEKLEQLQSHENEKIQKEAQEALEKQFSH
pI/MW: 5.09 / 27139.7

Mutant 4

MRGSHHHHHHGSELPQMTQQLNSDDMQEQLSATVKFRQILSRDG
NEQIQAVIDAGALPALVQLLSSPNEKLLKEAAWALSNIASGG
NEQIQAVIDAGALPALVQLLSSPNEKLLKEAAWALSNIASGG
NEQIQAVIDAGALPALVQLLSSPNEKLLKEAAWALSNIASGG
NEQIQAVIDAGALPALVQLLSSPNEKLLKEAAWALSNIASGG
NEQKQAVKEAGALEKLEQLQSHENEKIQKEAQEALEKQFSH
pI/MW: 5.09 / 26971.3

Mutant 5

MRGSHHHHHHGSELPQMTQQLNSDDMQEQLSATVKFRQILSRDG
NEQIQAVIDAGALPALVQLLSSPNEKLLKEAAWVLSNLASGG
NEQIQAVIDAGALPALVQLLSSPNEKLLKEAAWVLSNLASGG
NEQIQAVIDAGALPALVQLLSSPNEKLLKEAAWVLSNLASGG
NEQIQAVIDAGALPALVQLLSSPNEKLLKEAAWVLSNLASGG
NEQKQAVKEAGALEKLEQLQSHENEKIQKEAQEALEKQFSH
pI/MW: 5.09 / 27083.6

Mutant 6

MRGSHHHHHHGSELPQMTQQLNSDDMQEQLSATVKFRQILSRDG
NEQIQAVIDAGALPALVQLLSSPNEKLLKEAAWALSNLASGG
NEQIQAVIDAGALPALVQLLSSPNEKLLKEAAWALSNLASGG
NEQIQAVIDAGALPALVQLLSSPNEKLLKEAAWALSNLASGG
NEQIQAVIDAGALPALVQLLSSPNEKLLKEAAWALSNLASGG
NEQKQAVKEAGALEKLEQLQSHENEKIQKEAQEALEKQFSH
pI/MW: 5.09 / 26971.3

Mutant 7 (YM_4A (KK))
MRGSHHHHHHGSELPQMTQQLNSDDMQEQLSATVKFRQILSRDG
NEQIQAVIDAGALPALVQLLSSPNEKILKEALWALSNIASGG
NEQIQAVIDAGALPALVQLLSSPNEKILKEALWALSNIASGG
NEQIQAVIDAGALPALVQLLSSPNEKILKEALWALSNIASGG
NEQIQAVIDAGALPALVQLLSSPNEKILKEALWALSNIASGG
NEQKQAVKEAGALEKLEQLQSHENEKIQKEAQEALEKQFSH
pI/MW: 5.09 / 27139.7

Mutant 8
MRGSHHHHHHGSELPQMTQQLNSDDMQEQLSATVKFRQILSRDG
NEQIQAVIDAGAVPALVQLLSSPNEKLLKEALWALSNLASGG
NEQIQAVIDAGAVPALVQLLSSPNEKLLKEALWALSNLASGG
NEQIQAVIDAGAVPALVQLLSSPNEKLLKEALWALSNLASGG
NEQIQAVIDAGAVPALVQLLSSPNEKLLKEALWALSNLASGG
NEQKQAVKEAGALEKLEQLQSHENEKIQKEAQEALEKQFSH
pI/MW: 5.09 / 27083.6

Mutant 9
MRGSHHHHHHGSELPQMTQQLNSDDMQEQLSATVKFRQILSRDG
NEQIQAVIDAGAIPALVQLLSSPNEKLLKEALWALSNLASGG
NEQIQAVIDAGAIPALVQLLSSPNEKLLKEALWALSNLASGG
NEQIQAVIDAGAIPALVQLLSSPNEKLLKEALWALSNLASGG
NEQIQAVIDAGAIPALVQLLSSPNEKLLKEALWALSNLASGG
NEQKQAVKEAGALEKLEQLQSHENEKIQKEAQEALEKQFSH
pI/MW: 5.09 / 27139.7

Mutant 10
MRGSHHHHHHGSELPQMTQQLNSDDMQEQLSATVKFRQILSRDG
NEQIQAVIDAGALPALVQLLSSPNEKLLKEAVWALSNLASGG
NEQIQAVIDAGALPALVQLLSSPNEKLLKEAVWALSNLASGG
NEQIQAVIDAGALPALVQLLSSPNEKLLKEAVWALSNLASGG
NEQIQAVIDAGALPALVQLLSSPNEKLLKEAVWALSNLASGG
NEQKQAVKEAGALEKLEQLQSHENEKIQKEAQEALEKQFSH
pI/MW: 5.09 / 27083.6

Mutant 11
MRGSHHHHHHGSELPQMTQQLNSDDMQEQLSATVKFRQILSRDG
NEQIQAVIDAGALPALVQLLSSPNEKVLKEALWVLSNLASGG
NEQIQAVIDAGALPALVQLLSSPNEKVLKEALWVLSNLASGG
NEQIQAVIDAGALPALVQLLSSPNEKVLKEALWVLSNLASGG
NEQIQAVIDAGALPALVQLLSSPNEKVLKEALWVLSNLASGG
NEQKQAVKEAGALEKLEQLQSHENEKIQKEAQEALEKQFSH
pI/MW: 5.09 / 27195.8

Mutant 12
MRGSHHHHHHGSELPQMTQQLNSDDMQEQLSATVKFRQILSRDG
NEQIQAVIDAGALPALVQLLSSPNEKLLKEALWVLSNLASGG
NEQIQAVIDAGALPALVQLLSSPNEKLLKEALWVLSNLASGG
NEQIQAVIDAGALPALVQLLSSPNEKLLKEALWVLSNLASGG
NEQIQAVIDAGALPALVQLLSSPNEKLLKEALWVLSNLASGG
NEQKQAVKEAGALEKLEQLQSHENEKIQKEAQEALEKQFSH
pI/MW: 5.09 / 27251.9

Mutant 13
MRGSHHHHHHGSELPQMTQQLNSDDMQEQLSATVKFRQILSRDG
NEQIQAVIDAGAIPALVQLLSSPNEKLLKEALWVLSNLASGG
NEQIQAVIDAGAIPALVQLLSSPNEKLLKEALWVLSNLASGG

NEQIQAVIDAGAIPALVQLLSSPNEKLLKEALWVLSNLASGG
NEQIQAVIDAGAIPALVQLLSSPNEKLLKEALWVLSNLASGG
NEQKQAVKEAGALEKLEQLQSHENEKIQKEAQEALEKQFSH
pI/MW: 5.09 / 27251.9

Mutant 14
MRGSHHHHHHGSELPQMTQQLNSDDMQEQLSATVKFRQILSRDG
NEQIQAVIDAGALPALVQLLSSPNEKLLKEAVWVLSNLASGG
NEQIQAVIDAGALPALVQLLSSPNEKLLKEAVWVLSNLASGG
NEQIQAVIDAGALPALVQLLSSPNEKLLKEAVWVLSNLASGG
NEQIQAVIDAGALPALVQLLSSPNEKLLKEAVWVLSNLASGG
NEQKQAVKEAGALEKLEQLQSHENEKIQKEAQEALEKQFSH
pI/MW: 5.09 / 27195.8

Mutant 15
MRGSHHHHHHGSELPQMTQQLNSDDMQEQLSATVKFRQILSRDG
NEQIQAVIDAGALPALVQLLSSPNEKILKEAAWALSNLASGG
NEQIQAVIDAGALPALVQLLSSPNEKILKEAAWALSNLASGG
NEQIQAVIDAGALPALVQLLSSPNEKILKEAAWALSNLASGG
NEQIQAVIDAGALPALVQLLSSPNEKILKEAAWALSNLASGG
NEQKQAVKEAGALEKLEQLQSHENEKIQKEAQEALEKQFSH
pI/MW: 5.09 / 26971.3

Mutant 16
MRGSHHHHHHGSELPQMTQQLNSDDMQEQLSATVKFRQILSRDG
NEQIQAVIDAGGLPALVQLLSSPNEKLLKEALWALSNIASGG
NEQIQAVIDAGGLPALVQLLSSPNEKLLKEALWALSNIASGG
NEQIQAVIDAGGLPALVQLLSSPNEKLLKEALWALSNIASGG
NEQIQAVIDAGGLPALVQLLSSPNEKLLKEALWALSNIASGG
NEQKQAVKEAGALEKLEQLQSHENEKIQKEAQEALEKQFSH
pI/MW: 5.09 / 27083.6

Mutant 17
MRGSHHHHHHGSELPQMTQQLNSDDMQEQLSATVKFRQILSRDG
NEQIQAVIDAGAVPALVQLLSSPNEKLLKEALWVLSNIASGG
NEQIQAVIDAGAVPALVQLLSSPNEKLLKEALWVLSNIASGG
NEQIQAVIDAGAVPALVQLLSSPNEKLLKEALWVLSNIASGG
NEQIQAVIDAGAVPALVQLLSSPNEKLLKEALWVLSNIASGG
NEQKQAVKEAGALEKLEQLQSHENEKIQKEAQEALEKQFSH
pI/MW: 5.09 / 27195.8

Mutant 18
MRGSHHHHHHGSELPQMTQQLNSDDMQEQLSATVKFRQILSRDG
NEQIQAVIDAGAIPALVQLLSSPNEKILKEAAWALSNLASGG
NEQIQAVIDAGAIPALVQLLSSPNEKILKEAAWALSNLASGG
NEQIQAVIDAGAIPALVQLLSSPNEKILKEAAWALSNLASGG
NEQIQAVIDAGAIPALVQLLSSPNEKILKEAAWALSNLASGG
NEQKQAVKEAGALEKLEQLQSHENEKIQKEAQEALEKQFSH
pI/MW: 5.09 / 26971.3

Mutant 19
MRGSHHHHHHGSELPQMTQQLNSDDMQEQLSATVKFRQILSRDG
NEQIQAVIDAGGLPALVQLLSSPNEKLLKEAAWALSNLASGG
NEQIQAVIDAGGLPALVQLLSSPNEKLLKEAAWALSNLASGG
NEQIQAVIDAGGLPALVQLLSSPNEKLLKEAAWALSNLASGG
NEQIQAVIDAGGLPALVQLLSSPNEKLLKEAAWALSNLASGG
NEQKQAVKEAGALEKLEQLQSHENEKIQKEAQEALEKQFSH
pI/MW: 5.09 / 27915.2

YM_3A
The proteins contain 211 residues.
The extinction coefficient at 280 nm (ε) is 16500 M^{-1} cm^{-1} for all the proteins.

YM_3A
MRGSHHHHHHGSELPQMTQQLNSDDMQEQLSATVKFRQILSRDG
NEQIQAVIDAGALPALVQLLSSPNEKILKEALWALSNIASGG
NEQIQAVIDAGALPALVQLLSSPNEKILKEALWALSNIASGG
NEQIQAVIDAGALPALVQLLSSPNEKILKEALWALSNIASGG
NEQKQAVKEAGALEKLEQLQSHENEKIQKEAQEALEKQFSH
pI/MW: 5.23 / 22813.7

YM_4A (KK) variants
The proteins contain 253 residues.
The extinction coefficient at 280 nm (ε) is 22000 M^{-1} cm^{-1} for all the proteins.

$Y(KQ)_4A$
MRGSHHHHHHGSELPQMTQQLNSDDMQEQLSATVKFRQILSRDG
NEQIQAVIDAGALPALVQLLSSPNEKILQEALWALSNIASGG
NEQIQAVIDAGALPALVQLLSSPNEKILQEALWALSNIASGG
NEQIQAVIDAGALPALVQLLSSPNEKILQEALWALSNIASGG
NEQIQAVIDAGALPALVQLLSSPNEKILQEALWALSNIASGG
NEQKQAVKEAGALEKLEQLQSHENEKIQKEAQEALEKQFSH
pI/MW: 4.78 / 27139.5

$Y4(QK)_4A$
MRGSHHHHHHGSELPQMTQQLNSDDMQEQLSATVKFRQILSRDG
NEQIQAVIDAGALPALVQLLSSPNEQILKEALWALSNIASGG
NEQIQAVIDAGALPALVQLLSSPNEQILKEALWALSNIASGG
NEQIQAVIDAGALPALVQLLSSPNEQILKEALWALSNIASGG
NEQIQAVIDAGALPALVQLLSSPNEQILKEALWALSNIASGG
NEQKQAVKEAGALEKLEQLQSHENEKIQKEAQEALEKQFSH
pI/MW: 4.78 / 27139.5

$Y4(QQ)_4A$
MRGSHHHHHHGSELPQMTQQLNSDDMQEQLSATVKFRQILSRDG
NEQIQAVIDAGALPALVQLLSSPNEQILQEALWALSNIASGG
NEQIQAVIDAGALPALVQLLSSPNEQILQEALWALSNIASGG
NEQIQAVIDAGALPALVQLLSSPNEQILQEALWALSNIASGG
NEQIQAVIDAGALPALVQLLSSPNEQILQEALWALSNIASGG
NEQKQAVKEAGALEKLEQLQSHENEKIQKEAQEALEKQFSH
pI/MW: 4.53 / 27139.3

Unselected library members N3C
The proteins contain 211 residues.

1_3
MRGSHHHHHHGSELPQMTQQLNSDDMQEQLSATVKFRQILSRDG
NEQKQAVIDAGALPALVQLLSSPNEKILKTALEALENIAVVG
NEQKQAVIDAGALPALVQLLSSPNEKILKYALVALSNIAIKG
NEQRQAVIDAGALPALVQLLSSPNEKILKSALEALINIAIDG
NEQKQAVKEAGALEKLEQLQSHENEKIQKEAQEALEKQFSH
pI/MW: 5.62 / 22953.1
Extinction coefficient at 280 nm (ε): 1490 M^{-1} cm^{-1} (no Trp, more than 10% error in the value)

1_4
MRGSHHHHHHGSELPQMTQQLNSDDMQEQLSATVKFRQILSRDG
NEQIQAVIDAGALPALVQLLSSPNEKILKHALIALQNIAAFG
NEQKQAVIDAGALPALVQLLSSPNEKILKVALSALMNIADTG
NEQQQAVIDAGALPALVQLLSSPNEKILKFALEALNNIAHHG
NEQKQAVKEAGALEKLEQLQSHENEKIQKEAQEALEKQFSH
pI/MW: 5.73 / 22999.0
The proteins do not possess any Trp, Tyr or Cys and ε cannot be calculated.

2_1
MRGSHHHHHHGSELPQMTQQLNSDDMQEQLSATVKFRQILSRDG
NEQHQAVIDAGALPALVQLLSSPNEKILKMALWALSNIAAYG
NEQEQAVIDAGALPALVQLLSSPNEKILKAALIALRNIAKVG
NEQRQAVIDAGALPALVQLLSSPNEKILKMALWALWNIAEMG
NEQKQAVKEAGALEKLEQLQSHENEKIQKEAQEALEKQFSH
pI/MW: 5.84 / 23227.4
Extinction coefficient at 280 nm (ε): 17990 M^{-1} cm^{-1}

2_4
MRGSHHHHHHGSELPQMTQQLNSDDMQEQLSATVKFRQILSRDG
NEQEQAVIDAGALPALVQLLSSPNEKILKRALNALYNIAYNG
NEQQQAVIDAGALPALVQLLSSPNEKILKLALDALVNIAANG
NEQHQAVIDAGALPALVQLLSSPNEKILKDALRALINIASSG
NEQKQAVKEAGALEKLEQLQSHENEKIQKEAQEALEKQFSH
pI/MW: 5.56 / 23010.9
Extinction coefficient at 280 nm (ε): 2980 M^{-1} cm^{-1} (no Trp, more than 10% error in the value)

2_7
MRGSHHHHHHGSELPQMTQQLNSDDMQEQLSATVKFRQILSRDG
NEQQQAVIDAGALPALVQLLSSPNEKILKIALEALHNIAVMG
NEQQQAVIDAGALPALVQLLSSPNEKILKHALRALSNIADDG
NEQHQAVIDAGALPALVQLLSSPNEKILKFALNALWNIAWHG
NEQKQAVKEAGALEKLEQLQSHENEKIQKEAQEALEKQFSH
pI/MW: 5.70 / 23219.2
Extinction coefficient at 280 nm (ε): 11000 M^{-1} cm^{-1}

2_8
MRGSHHHHHHGSELPQMTQQLNSDDMQEQLSATVKFRQILSRDG
NEQIQAVIDAGALPALVQLLSSPNEKILKYALWALINIAFNG
NEQHQAVIDAGALPALVQLLSSPNEKILKDALQALSNIAEYG
NEQHQAVIDAGALPALVQLLSSPNEKILKHALAALRNIAWQG
NEQKQAVKEAGALEKLEQLQSHENEKIQKEAQEALEKQFSH
pI/MW: 5.73 / 23247.2
Extinction coefficient at 280 nm (ε): 13980 M^{-1} cm^{-1}

2_9
MRGSHHHHHHGSELPQMTQQLNSDDMQEQLSATVKFRQILSRDG
NEQRQAVIDAGALPALVQLLSSPNEKILKVALDALMNIALMG
NEQTQAVIDAGALPALVQLLSSPNEKILKHALYALWNIAYRG
NEQQQAVIDAGALPALVQLLSSPNEKILKAALTALYNIAWRG
NEQKQAVKEAGALEKLEQLQSHENEKIQKEAQEALEKQFSH
pI/MW: 5.98 / 23293.4
Extinction coefficient at 280 nm (ε): 15470 M^{-1} cm^{-1}

2_10
MRGSHHHHHHGSELPQMTQQLNSDDMQEQLSATVKFRQILSRDG
NEQIQAVIDAGALPALVQLLSSPNEKILKHALIALINIAMWG
NEQEQAVIDAGALPALVQLLSSPNEKILKKALFALANIAYKG
NEQTQAVIDAGALPALVQLLSSPNEKILKYALEALQNIAFHG
NEQKQAVKEAGALEKLEQLQSHENEKIQKEAQEALEKQFSH
pI/MW: 5.78 / 23201.4
Extinction coefficient at 280 nm (ε): 8480 M^{-1} cm^{-1}

2_12
MRGSHHHHHHGSELPQMTQQLNSDDMQEQLSATVKFRQILSRDG
NEQIQAVIDAGALPALVQLLSSPNEKILKIALHALVNIAKSG
NEQRQAVIDAGALPALVQLLSSPNEKILKFALDALSNIASWG
NEQQQAVIDAGALPALVQLLSSPNEKILKDALNALFNIAKTG
NEQKQAVKEAGALEKLEQLQSHENEKIQKEAQEALEKQFSH
pI/MW: 5.83 / 23025.0
Extinction coefficient at 280 nm (ε): 5500 M^{-1} cm^{-1}

2_14
MRGSHHHHHHGSELPQMTQQLNSDDMQEQLSATVKFRQILSRDG
NEQRQAVIDAGALPALVQLLSSPNEKILKSALTALRNIASSG
NEQHQAVIDAGALPALVQLLSSPNEKILKWALEALTNIAEEG
NEQEQAVIDAGALPALVQLLSSPNEKILKKALNALMNIAVHG
NEQKQAVKEAGALEKLEQLQSHENEKIQKEAQEALEKQFSH
pI/MW: 5.67 / 23060.0
Extinction coefficient at 280 nm (ε): 5500 M^{-1} cm^{-1}

Acknowledgments

A Ph.D. thesis is the result of several years of work but also of the interaction with many people who made practical contributions or just suggestions that I tried to follow.

First of all, I want to thank Andreas Plückthun for the opportunity he offered me with this project and the freedom to shape it as I wanted, believing in my ideas even when they were slightly different from his. Thanks also to the members of my Ph.D. committee, Donald Hilvert and Markus Grütter for their support and to Oliver Zerbe for the NMR measurements and his help.

I want to acknowledge who contributed directly to this project. Michael Stumpp for his “supervision”, especially during my first period in the group. His help and suggestions have been critical to start in the right direction without getting lost. Anders Peter Larsen for being my student and teaching me how to teach. I probably learned more then he did in the year we spent together in the lab. Kai Hoch for donating his hands and his time in Zürich as student to the project. Annemarie for her macros and analysis. Sarel Fleishman for a long distance and enthusiastic collaboration. Riccardo Pellarin and Pietro Alfarano for their help and significant contribution, bringing computation, molecular dynamics and new points of view. As well as friendship. Gautham Varadamsetty, Dirk Tremmel and Christian Reichen for joining this project, sharing work, ideas and problems. It is satisfactory to know that somebody will continue what I started.

Thanks to the ankyrin people (most of them now in Molecular Partners) and in particular Kaspar Binz for his design work: without it probably I would not have decided to start this project.

Thanks to the lab members for their discussions and help, but especially for turning my time in the lab (and outside) into a nice life much more than just a job. I will never thank you enough for this. I want to thank in particular Jonas, who shared a lab for two with me for some years and tells me that he could even stand it for longer, and Myriam, who shared with me Ph.D. applications, groups, lunches and dinners and more and surprisingly she did not have enough. Thanks also to my almost acquired group (the Caflisches), for all the interactions, discussions and collaborations we had.

An acknowledgment goes to all the people that made possible my work, in term of bureaucracy, infrastructure and materials: Peter Lindner and Petra Vogt first, but also all the members of the biochemistry staff, from IT to logistic, secretariat, workshop, maintenance and cleaning. And thanks to all the websites from which I stole all the armadillo pictures that I used in the last years.

I want to thank also all the people that helped and supported me, enriching my life in the last years. My parents and relatives for their trust and support. All the new friends I met from all countries for the nice time we shared. The members of the Italian community of Zürich, for the life outside the lab, their support and their Italian. The people of the MLS program and in particular the few that shared much more than just lectures, retreats and parties. My old friends from Italy (RPG players and not) that stayed in touch. My flatmates of Frohburgstrasse 198, the former and especially the last ones who became my sort of family. Flavio for his spirit, his music, his dinners, his friendship and sharing all or them. Schu-fee for being the best German/Taiwanese friend I ever met, despite her tastes concerning certain types of food.

Thanks to all the people above who tried, unsuccessfully, to teach me the importance of being on time. Do not be disappointed, nobody managed so far, even if I started to think about it. Maybe it is the age.

Looking backwards I can say that I had Luck during the last years. And probably I should also acknowledge it, because you always need it. In science and life.

Curriculum vitae

First Name: Fabio
Last Name: Parmeggiani
Nationality: italian
Birthdate: April 4th 1978
Birthplace: Ferrara, Italy

Education
2004–2008: Ph.D. position in the group of Prof. A. Plückthun, Department of Biochemistry, University of Zürich.
2006: Biosensor workshop on SPR technology with Prof. D. G. Myszka**,** Salt Lake City (USA).
2004-2005: research fellowship from Roche Research Foundation.
2003: academic guest at the Institute for polymers, ETH Zürich, Switzerland (group of Prof. P.L. Luisi).
2003: Master in Biotechnology, University of Bologna, with a final examination mark of 110/110 *cum laude*.
2002: master's work in the laboratory of Prof. A. Hochkoeppler, Department of Industrial Chemistry, University of Bologna.
2001: summer student at EMBL, Heidelberg, Germany (A. Hoenger group).
1999: english course in London, UK.
1997 – 2003: university studies in Biotechnology at the University of Bologna, Italy.
1997: italian high school diploma (liceo scientifico A. Roiti of Ferrara) with a final examination mark of 60/60.

Publications
Parmeggiani F., Pellarin R., Larsen A.P., Varadamsetty G., Stumpp M.T., Zerbe O., Caflisch A., Plückthun A.
"Designed Armadillo Repeat Proteins as general peptide binding scaffolds"
J. Mol. Biol. (2008) 376, 1282–1304.

Navratilova I., Papalia G. A., Rich R. L., Bedinger D., Brophy S., Condon B., Deng T., Emerick A. W., Guan H. W., Hayden T., Heutmekers T., Hoorelbeke B., McCroskey M. C., Murphy M. M., Nakagawa T., Parmeggiani F., Qin X., Rebe S., Tomasevic N., Tsang T., Waddell M. B., Zhang F. F., Leavitt S., Myszka D. G.
"Thermodynamic benchmark study using Biacore technology".
Anal. Biochem. (2007) 364, 67-77.

Radeghieri A., Bonoli M., Parmeggiani F., Hochkoeppler A.
"Tyrosine83 is essential for the activity of *E. coli* galactoside transacetylase".
Biochim. Biophys. Acta (2007) 1774, 243-8.

Parmeggiani F., Pellarin R., Larsen A.P., Varadamsetty G., Stumpp M.T., Plückthun A.
Patent application EP07117059.1 "Designed Armadillo Repeat Proteins" (2007).